nature portfolio

Peer Review File

Dopamine Release Plateau and Outcome Signals in Dorsal Striatum Contrast with Classic Reinforcement Learning Formulations

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This is a technically advanced, well-designed and readily understandable report on experiments designed to measure the changes in extracellular dopamine in the medial and lateral parts of the striatum during acquisition and performance of reinforcement learning tasks. The paper is very clearly written and the figures are do a good job of illustrating the experiments and the findings.

The fundamental finding is a failure to confirm the predictions of the reward prediction error role for dopamine in reinforcement learning. In this formulation, transient dopamine increases in the striatum act as an instructive signal that is evoked by rewarding stimuli. During Pavlovian associative learning shifts earlier in the sequence of events that predict the reward, finally resting on the earliest cue that predicts a sequence that ends with a reward. This has been a very influential idea, both in the neurobiology of learning and in robotics and psychology. The idea was based on studies of changing dopaminergic neuron firing rates during learning, and since its inception, the range of cellular mechanisms known to control dopamine concentration in the striatum has expanded, and so it is valuable to measure dopamine, not simply dopaminergic neuron activity. There are now very powerful methods for measuring extracellular dopamine activity with good spatial and temporal resolution, and this paper uses one of those.

In its simplest and most often stated form, the RPE model states that dopamine controls synaptic plasticity (and the learning rate) to enhance synaptic weights for inputs that predict reinforcement. It is expected that dopamine release will occur in response to a primary reward, and not (or much less) to a neutral cue. When the neutral cue becomes reward-predicting, the dopamine response should shift earlier in time during learning and become associated with early cues according to their predictive value. The response to the reward should decrement as this is happening. The authors report that the response to the cue does not increase during acquisition and the response to the reward does not decrement. In addition, in the simple form of RPE, dopamine levels should have the same reinforcement learning function in all parts of the striatum. The authors reported that responses to reward were often opposite in sign in medial and lateral parts of the striatum. A comparison with the larger literature on dopamine in the ventral striatum reveals further heterogeneity.

This is a well-executed study that provides strong evidence against the RPE model in the dorsal striatum, but it leaves me wondering where we go from here. The dopamine system was one of the best examples of a set of neurons whose firing seemed to correspond directly to a psychological state. This is the reason for the neuroscientists' and psychologists' obsession with it. I read this paper as concluding that it just isn't so – the dopamine signal is actually as multidimensional and undecipherable as all the other neuronal signals that have been studied in detail. If this is so, it contradicts the whole point of measuring dopamine as a single thing with a single meaning. Dopamine's meaning in learning would be different on different neurons, or even different synapses. It would be like measuring extracellular glutamate in the cortex. At least it casts doubt on the role of dopamine as an instructional signal in learning. Perhaps we should look elsewhere for that. I realize that Nature Communications is for short papers full of data, and not a place to reflect on their implications, but the implications of this work seem too important (and too destructive of current consensus) to go unaddressed.

Reviewer #2 (Remarks to the Author):

This manuscript by Kim et al. investigates dopamine release encoding during many cue conditioning tasks in medial and lateral parts of the dorsal striatum, specifically in relation to RPE model. As they stated in the introduction, whether RPE can explain dopamine release has been controversial and actively tested especially in the ventral striatum. Thus, the motivation of the study is solid, and the amount of data they collected from different parts of dorsal striatum with careful single fiber insertions across many mice in six different tasks is impressive. Overall, I think the manuscript makes a strong contribution to the field. That said, I have some concerns primarily resulting from the descriptive nature of the paper. As a reader, I thought it was hard to arrive at a concrete conclusion due to scattered results and descriptions. This can be improved with clearer interpretation of the data and experimental design.

Major

1. A key argument made by the authors against the RPE model is the emergence of plateau responses during discrimination learning. The authors present this as self-evidently against RPE. However, this is not clear to me as a reader. Prolonged activity (e.g., ramping) can exist during intermediate time between cue and reward and has previously been argued to be consistent with RPE (e.g., Kim et al. 2020). Why do authors think that the prolonged response is against RPE? Is this because they first see the strong cue onset activity in the earlier sessions of cue discrimination task, and then started to observe prolonged activity in the later sessions? Or did this prolonged activity arise first from cue onset and then elongate toward reward, which is in the opposite direction of ‘backward shifting’? It is hard to be convinced without a clear rationale. In fact, with sensory uncertainty related arguments as have been made previously in Kim et al. 2020 and Mikhael et al. 2021, I could imagine obtaining these results using RPE encoding. If the authors can simulate a TDRL model and show that this is inconsistent with RPE encoding, that would be more convincing.

2. Related to the above point, cue discrimination and single-cue sessions appear to have different reward and cue rates. It’s unclear if the emergence of a plateau response has to do with the presence of CS- trials as argued by the authors, or concomitant changes in reward/cue rate in the task. Of course, addressing this would require additional data collection, which may not be easily

feasible. So, I’m okay with the authors discussing that the reason for the observed prolonged cue response remains unclear.

3. The authors show that the plateau response is not a result of averaging across mice, but it would be useful to also test if it is a result of averaging across trials within individual mice.

4. Along similar lines, it is unclear if the observed effects might reflect a strong order effect of the exact experimental conditions that were run. Every animal went through the same sequence of history and thus, it is hard to be sure that the observations here reflect task differences as claimed by the authors, or experimental order effects. Some counterbalancing and controls would be necessary to better interpret the results. Again, since this is a lot of work, I am okay with the authors just discussing this weakness. That said, the methodological details in #2 and in this point diminish my enthusiasm for the strength of the data.

5. Though the authors contrast the results observed here with Amo et al. 2022, the claim in this paper is not just that there is no systematic backpropagation during the delay period from reward to cue (like Amo et al. 2022) but that the cue response does not increase at all. Since the lack of apparent backpropagation during intermediate timepoints between cue and delay can be explained by a high eligibility trace (lambda), but the current findings cannot be explained by a high lambda, the authors may want to highlight this lack of increasing cue response as a stronger test of reward-to-cue response shift than that considered in prior studies. Related to this, Jeong et al. also claimed to not find the backpropagating bump that Amo et al. found. This may be worth mentioning for completeness. However, this study too was claiming that the cue response increased over learning. Thus, the strength of the observations here is somewhat muddled by the description. I would encourage the authors to highlight this point more.

6. One point that I found quite interesting in Jeong et al. 2022 was the increase of dopamine responses to random rewards. They argued that this is inconsistent with RPE since the initial presentations of rewards are the most unpredicted. The authors similarly find here that random reward responses increase across training. This appears to also be in violation of TD-RPE. The authors may want to highlight this finding. Based on my understanding, an RPE (received minus predicted reward) signal can increase only if prediction is negative or received reward changes. Neither appears to be the case here, suggesting that this is an interesting result to highlight.

7. To be honest, I felt that it’s a missed opportunity that the authors did not discuss (i.e., Discussion section) the difference between medial vs. lateral striatum. I think these data has great potential to discuss this matter, instead of just focusing on RPE. On a related note, I thought Figure 3e is very interesting. Even in non-learners, there was a clear distinction in dopamine release to cues in DLS, but not in DMS. Is this related to the motor control aspect of DMS?

8. Related to the above finding, it is already known that DLS has higher DAT expression and hence, faster dynamics than DMS. Can this instruct the conclusion in Figure 5 (e.g., DMS can appear to have more plateau-like activity because of its slow timescale)?

Minor

1. In the discussion, the authors describe the Jeong et al. study and the current study as measuring from dopamine axons. Both studies use dopamine release measurements and not dopamine axon GCaMP dynamics. It may be better to be more careful about this description.

2. The authors state that “The recordings made by Jeong et al., like our recordings, were made from dopamine-containing axons, not from dopamine cell bodies as in the original studies linking dopamine dynamics to RPE. These discrepancies could be accounted for according the Uchida and Watabe-Uchida and Uchida groups (e.g., Ref 40; see also Ref 39).” This sentence seems a bit unclear to me. Are the authors stating that Ref 40 explains the discrepancies between RPE and dopamine release data from Jeong et al., or that Ref 40 found no discrepancy between cell body activity and dopamine axon/release dynamics? I think they might potentially mean the latter, but the sentence reads like the former. Since the Jeong et al. paper is more recent than Ref 40, I am confused about how Ref 40 accounts for the Jeong et al. results. Perhaps I am misreading this sentence but it is good to be clear.

Reviewer #3 (Remarks to the Author):

Several recent studies have challenged the classical hypothesis that midbrain dopamine (DA) neurons signal reward prediction error (RPE) to the striatum. The authors tested this issue for the dorsal striatum by training mice consecutively on a series of visual cue-outcome conditioning tasks and recording dopamine release responses with chemosensor probes throughout the time that the mice were learning the tasks. The Introduction beautifully summarizes the rationale of the aims. Experiments are in most parts very well thought out.

Recording sites are mostly within a small anterior dorsal striatal subregion.

The results clearly show that DA release profiles in the dorsal striatum undergo very dynamic learning-related changes. An amazing finding is that within the small region DA signals show clear location dependent differences which are consistent among animals. These different release profiles were also selective for different versions of associative cue-outcome conditioning. Importantly, they found no evidence for a transition from outcome to cue signaling that a hallmark of temporal difference reinforcement learning as applied to the substantia nigra pars compacta (SNpc) activity reported in previous studies. Results greatly extend recent concept that SNpc DA neurons and the striatum have heterogeneous and specialized responses to task variables during complex behavior. I totally agree with the authors suggestion that “further refinement and extension of reinforcement learning algorithms is needed to account for spatiotemporal DA release dynamics in the dorsal striatum”.

The results are noteworthy and significant for the advancement of the basal ganglia research. Comments below are to clarify some minor issues in the manuscript.

The limitations of the studies are clearly described. One such limitation is that possible stimulation of DA fibers by acetylcholine in the striatum in vivo is speculative. Two very recent Nature articles (s41586-023-06492-9_reference and s41586-023-05995-9) describing acetylcholine and DA dynamics in the striatum are relevant to this issue.

Reason to target two sites within the small dorsal striatum become clear in later part of the manuscript. It could be mentioned in the Method or in Introduction why the medial site (AP: +1.0 mm, ML: +1.5 mm, DV: −2.7 mm from bregma) and lateral site (AP: +1.0 mm, ML: +1.9 mm, DV: −3.0 mm from bregma) which have only 0.5mm separation were targeted. Also, please describe the diameter and the shape of optic probe tip.

Most of recording sites shown in F5 and supplemental 1 are in the medial sector of the striatum, if “Recording sites were classified as "medial" if the distance from the midline to the tip of the probe was less than 0.6 of the distance from the midline to the lateral edge of the striatum.”

Fig. 1b, please give the size of scale bar.

Fig. 1b and 1F, DMs are in mm. Please describe the location of the reference point (DV=0.0).

Fig. 1f, Mixed use of mm and standardized coordinates for probe locations is confusing.

Methods to standardize coordinates for ML and DV locations are described but not for AP used in Fig. 5 and supplemental 1.

Reviewer #4 (Remarks to the Author):

Kim et al. recorded dopamine sensor signals with a single fiber in the dorsomedial striatum (DMS) or dorsolateral striatum (DLS) in head-fixed mice. The mice experienced a series of tasks with water reward, some with visual cues, a light on the left or right. Authors found great individual variability in the sensor signals, some explained by fiber locations, sexes and learning extents. From these complicated results, authors concluded that dopamine release in the dorsal striatum is not consistent with reward prediction error.

Since signature studies in monkeys by Montagues et al., 1996 and Shultz et al., 1997, dopamine neuron activity had often been interpreted as reward prediction errors. However, many recent rodent studies as well as some monkey studies discovered that dopamine activity is more complicated than it had been believed, and that dopamine neurons are diverse at molecular levels as well as at functional levels, some of which correspond to diversity of projection targets. Even though the precise interpretation is different among studies, there is a consensus now that dopamine neurons are not uniform. This study by Kim et al. adds another example to yield such a conclusion. Although the presented results are potentially interesting, most findings are overlapped with previous studies. Most importantly, main findings should be verified with proper controls and analyses. There are additional concerns about data quality and data exclusion.

Major concerns (not exhaustive)

1. After several different tasks, dopamine release both in DMS and DLS became consistent with RPE. This was interpreted by authors that the later tasks are more cognitively demanding. Although this is an interesting idea, it is similarly possible that animals just needed to be trained longer in each task for dopamine neurons to exhibit RPE-like activity patterns. Multiple studies show that simple behavioral changes such as anticipatory licking are observed even before visible changes in population activity of dopamine neurons (ref. Menegas et al., 2017). Authors should have controls to train longer without switching tasks to determine whether cognitive demanding affects RPE-like signals in dopamine.

2. Single animals were used to test dopamine activity in multiple tasks without considering any effects of history. For example, mice were intensively trained with Operant tasks (lick to trigger water delivery) with almost fixed inter-trial-interval (ITI 6-8s) before starting random reward sessions. This is a quite complicated situation for animals; they are trained to expect water every time they lick around 6-8s after the previous water, and then now such a lick is not valid anymore, but instead they have to wait longer (ITI 8-48s) for passive delivery. Indeed, the lick trace in Figure 1d columns 1&2 show that what the animal learned was to lick less during ITI. While authors indicate that their observation in dopamine activity is not consistent with TD error, it is not clear what kind of TD learning should explain this learning and what kind of TD errors are expected to explain dopamine activity in such a complicated situation, switching between Operant and passive behaviors.

Similarly, those animals were then trained to associate a single cue and water, after intensive random water sessions. Again, under the history of high expectation of random water, introduction of a cue in this task trains animals to hold licking during ITI, rather than to make them to lick more after a cue (Figure 1d column 3&4). What is authors' prediction for TD errors in this task?

In the next task, another cue for no outcome was introduced, after animals already learned a reward-predicting cue. Why, then, should water responses in dopamine in this task be moved from outcome to a reward-predicting cue again as authors predicted from TD errors? The example animal in Figure 1d column 5&6 even learned to suppress licking after a reward-predicting cue, instead of increasing anticipatory licking.

Finally, authors observed RPE-like dopamine signals in later tasks in reversal and probabilistic tasks (i.e. decrease of reward responses and increase of cue responses over training and inhibition by reward omission, monotonic activation modulation by reward probability), with some differences between DMS and DLS, which is consistent with a previous study (Parker et al., 2016, Tsutsui-Kimura et a., 2020) and contradicts with authors' main claim.

3. Related to 2, while authors contrast their observation with TD errors, there is no formal prediction or no formal test. For example, if they want to claim for (or against) transfer of signals from outcome periods to cue periods, they should quantify the time-course of those signals at the trial basis in single animals. Because of complication of history until cue discrimination (see above), the reversal task would be a proper task for such quantification in this dataset. Basically, any claim should be based on proper quantification, which is largely lacking in the manuscript.

4. "Plateau response" at cue is sustained only for 1.5s, since there is no delay from cue and outcome in this task. Considering these slow dopamine sensors, this activity is likely the second peak that previous studies observed in dopamine neurons: first detection, and then value (Nomoto et al., 2010). Authors should have a long delay period from a cue to an outcome or examine electric spikes to determine whether the activity plateaus. In addition, PCA is indirect to quantify the sustained activity. The activity level during later periods should be directly used.

5. Unique responses to visual stimuli in dopamine in DMS had been reported (Moss et al., 2021). Thus, the initial quick cue responses are likely to be sensory (or detection), and may not be subject to value transfer, but instead may decay with familiarity. Authors should have controls without any outcomes (single cue, no outcome).

6. Previous studies reported contra-lateral bias in dopamine signals, especially in DMS (Moss et al, 2021). Responses to visual cues contra- vs ipsi-lateral to the recording site, and to associated outcomes should be analyzed separately.

7. Inhibition of dopamine activity by licks in initial stages of training had been reported with specific interpretation (Coddington and Dudman, 2018). Authors should test the original authors' model to be a follow-up study.

8. Difference of activity patterns between DMS and DLS should be quantified for each claim. Because sex and performance dramatically affected dopamine activity in this study, authors should consider those together instead of testing one-by-one, to determine contribution of location.

9. Half of sessions in the probabilistic reward task were excluded and "the session with the better recording quality was selected for analysis". Why do not they just average both sessions? How was the "recording quality" in other tasks?

More comments (not minor)

10. Why are signals even before a cue always synchronized on different days or in different locations such as in Figure 2a column2 and Figure 3a, c, d, e?

11. Related to 2, dynamical changes of lick patterns (during ITI and cue periods) in each trial type in each task across days should be shown to verify proper learning.

12. Learning progress should be summarized to show how many animals progressed to the next level and how long animals spent for each stage of learning.

13. Major findings should be verified using GFP expression (or GRAB-DA-mut, see Costa et al., 2023) as controls for motion artifacts in each area with the same normalization with 405w light.

14. Average activity patterns in all the figures should have error bars, instead of having only averages.

15. Z-score should be calculated using only ITI to compare activity across sessions. Show all trials in an example animal to verify success of normalization.

16. Task structures such as ITI distribution (uniform vs exponential) and trial types (% free water) should be clearly written.

17. Are Figure 2b-d from lateral or medial DS? Both should be shown separately.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This is a technically advanced, well-designed and readily understandable report on experiments designed to measure the changes in extracellular dopamine in the medial and lateral parts of the striatum during acquisition and performance of reinforcement learning tasks. The paper is very clearly written and the figures are do a good job of illustrating the experiments and the findings.

We thank the Reviewer for these positive and encouraging remarks.

The fundamental finding is a failure to confirm the predictions of the reward prediction error role for dopamine in reinforcement learning. In this formulation, transient dopamine increases in the striatum act as an instructive signal that is evoked by rewarding stimuli. During Pavlovian associative learning shifts earlier in the sequence of events that predict the reward, finally resting on the earliest cue that predicts a sequence that ends with a reward. This has been a very influential idea, both in the neurobiology of learning and in robotics and psychology. The idea was based on studies of changing dopaminergic neuron firing rates during learning, and since its inception, the range of cellular mechanisms known to control dopamine concentration in the striatum has expanded, and so it is valuable to measure dopamine, not simply dopaminergic neuron activity. There are now very powerful methods for measuring extracellular dopamine activity with good spatial and temporal resolution, and this paper uses one of those.

In its simplest and most often stated form, the RPE model states that dopamine controls synaptic plasticity (and the learning rate) to enhance synaptic weights for inputs that predict reinforcement. It is expected that dopamine release will occur in response to a primary reward, and not (or much less) to a neutral cue. When the neutral cue becomes reward-predicting, the dopamine response should shift earlier in time during learning and become associated with early cues according to their predictive value. The response to the reward should decrement as this is happening. The authors report that the response to the cue does not increase during acquisition and the response to the reward does not decrement. In addition, in the simple form of RPE, dopamine levels should have the same reinforcement learning function in all parts of the striatum. The authors reported that responses to reward were often opposite in sign in medial and lateral parts of the striatum. A comparison with the larger literature on dopamine in the ventral striatum reveals further heterogeneity.

This is a well-executed study that provides strong evidence against the RPE model in the dorsal striatum, but it leaves me wondering where we go from here.

We greatly appreciate this comment. This comment highlights what we and others are increasingly finding—namely, that the striatum is a structure with a large range of attributes, distributed across space and time. Even not very distant sites can have different afferent and/or efferent connections, different mixes of neurotransmitters, receptors, modulators and their affiliated molecules, local network regulators, and functions. This is not, however, a reason to despair; the field will figure out more about how the striatum is organized, and our prediction is that much of this will then make sense. We could consider the striatum in relation to the neocortex: no one worries that there are many different functionally or anatomically defined ‘areas’ and that, within them, their resident neurons exhibit diverse functional properties. The Reviewer’s point is an excellent one; we need to think more deeply, with more powerful methods and models. We have emphasized this in the last paragraph in our revised discussion section.

> '**Summary and caveats related to the findings**
> Here, we have shown discrepancies in both space and time between dopamine release patterns and patterns predicted by RPE formulations. These results corroborate the idea that the striatum is a composite of zones participating in multiple functional circuits. Cells involved in these circuits compute information in unique ways not necessarily equivalent to those of RPE formulations. Striatal microcircuitry is complex and spatially heterogeneous. It is possible that all or many regions of the striatum perform a similar core computation, but that single regions deal with different input-output and local circuit modulation according to requirements of given contexts and circumstances. Detailed study of the full range of variation in striatal dopamine response profiles could help to uncover the remarkable functional range of dopamine-based systems in modulating adaptive behavior'

The dopamine system was one of the best examples of a set of neurons whose firing seemed to correspond directly to a psychological state. This is the reason for the neuroscientists' and psychologists' obsession with it. I read this paper as concluding that it just isn't so – the dopamine signal is actually as multidimensional and undecipherable as all the other neuronal signals that have been studied in detail. If this is so, it contradicts the whole point of measuring dopamine as a single thing with a single meaning.

The Reviewer puts this so well; we cannot look for a single meaning. Moreover, we must be aware of the many local intrastriatal effects that can take a given dopamine signal and modify it. There are local control mechanisms (some of which we mention) and interactions with other neuromodulators (probably only partially recognized so far). Further, it is increasingly being shown that the dopamine-containing neurons of the substantia nigra themselves have different signaling properties, even some not overtly related to reward or reinforcement.

> '**Summary and caveats related to the findings**
> Here, we have shown discrepancies in both space and time between dopamine release patterns and patterns predicted by RPE formulations. These results corroborate the idea that the striatum is a composite of zones participating in multiple functional circuits. Cells involved in these circuits compute information in unique ways not necessarily equivalent to those of RPE formulations. Striatal microcircuitry is complex and spatially heterogeneous. It is possible that all or many regions of the striatum perform a similar core computation, but that single regions deal with different input-output and local circuit modulation according to requirements of given contexts and circumstances. Detailed study of the full range of variation in striatal dopamine response profiles could help to uncover the remarkable functional range of dopamine-based systems in modulating adaptive behavior'

Dopamine's meaning in learning would be different on different neurons, or even different synapses. It would be like measuring extracellular glutamate in the cortex. At least it casts doubt on the role of dopamine as an instructional signal in learning. Perhaps we should look elsewhere for that. I realize that Nature Communications is for short papers full of data, and not a place to reflect on their implications, but the implications of this work seem too important (and too destructive of current consensus) to go unaddressed.

We thank the Reviewer for this urging. We now have added text in the Discussion section to address these issues.

> 'We are aware of caveats that should accompany our conclusions. The tasks were variants of Pavlovian tasks and lacked the richness of much behavioral learning, decision-making and response variety. We used a fixed sequence of paradigms across animals as representative of the many switches that can occur in daily experience, but we are aware that the results could be constrained by this training sequence. We used both D1R-based (i.e., dLight1.3b and $\mathrm{GRAB}_{\mathrm{DA3m}}$) and D2R-based (i.e., $\mathrm{GRAB}_{\mathrm{DA2m}}$) dopamine sensors. Decay time constants for $\mathrm{GRAB}_{\mathrm{DA2m}}$ and $\mathrm{GRAB}_{\mathrm{DA3m}}$ are, respectively, 1.3 sec[48] and ~600 msec[66], but that for dLight1.3b has not been determined. This imposes a lower temporal resolution on our data as compared to electrical recordings. We only sampled relatively restricted parts of more medial and lateral parts of the centrodorsal striatum, and did not consider the compartmentalization of the striatum, in which striosome and matrix compartments have different relationships to dopamine[33,35,67]. We used photometry, a recording method that measures the local sum of extracellular dopamine, whereas dopamine likely works both at individual synapses[68] and as an ambient non-synaptic modulator. Our findings cannot address the synaptic actions of dopamine because our measurements are probably dominated by extrasynaptic dopamine. RPE-observing dopamine signaling might instruct reinforcement plasticity only in a small subset of synapses that convey the relevant information, as suggested by reports of multiple, multiplexed responses of dopamine and dopamine neuron firing[10,43,44,59-61]. Despite these uncertainties, the surprises that emerged in our experiments open new opportunities to probe and to model mechanisms underlying striatum-based learning and its modulation by dopamine.'

Reviewer #2 (Remarks to the Author):

This manuscript by Kim et al. investigates dopamine release encoding during many cue conditioning tasks in medial and lateral parts of the dorsal striatum, specifically in relation to RPE model. As they stated in the introduction, whether RPE can explain dopamine release has been controversial and actively tested especially in the ventral striatum. Thus, the motivation of the study is solid, and the amount of data they collected from different parts of dorsal striatum with careful single fiber insertions across many mice in six different tasks is impressive. Overall, I think the manuscript makes a strong contribution to the field. That said, I have some concerns primarily resulting from the descriptive nature of the paper. As a reader, I thought it was hard to arrive at a concrete conclusion due to scattered results and descriptions. This can be improved with clearer interpretation of the data and experimental design.

We are extremely sorry not to have arranged the results and descriptions with sufficient order. In our revision, we have tried to lay out the organization and the motivation for the tests and findings more systematically, and we have added subheadings to the text. We should have done this from the start. A great thanks to the Reviewer for this admonition. We have also added this initial paragraph to the Discussion:

> 'Our experiments with simple Pavlovian tasks lead to three major findings that suggest the need to review current RL-RPE models of dopamine's functions in the striatum. First, in the centromedial striatum, dopamine exhibited no or negative reward-associated outcome responses, in contrast to what was expected based on the RPE interpretation of dopamine activity. Also, in both centromedial and centrolateral striatum, the dopamine response to random reward increased with training, whereas an RPE signal would decrease with training. The RPE model is thus not sufficient to account for these data. Second, phasic dopamine release responses did occur to both the conditioned cue and to reward outcome in the centrolateral striatum, but with training the outcome response did not decline, and

> the cue response did not increase, also in contrast to the RPE interpretation of dopamine activity. Third, with discrimination learning, plateau-like responses, which tended to bridge the cue and reward associated responses, emerged and were strongest in the best performers, but almost absent in non-learners. Simple RL models, prima facie, do not predict the emergence of such responses, though with complex RL models they might appear (see below). We conclude that, at least at the population level that can be imaged by fiber photometry, dorsal striatal dopamine release responses do not fully follow RPE formulations in either more medial or more lateral regions, but exhibit instead unpredicted heterogeneities across striatal districts. These findings encourage further work on how the multitudes of striatal circuits are coordinated to instruct learning and to modulate behavior under the influence of dopamine.'

Major

1. A key argument made by the authors against the RPE model is the emergence of plateau responses during discrimination learning. The authors present this as self-evidently against RPE. However, this is not clear to me as a reader. Prolonged activity (e.g., ramping) can exist during intermediate time between cue and reward and has previously been argued to be consistent with RPE (e.g., Kim et al. 2020). Why do authors think that the prolonged response is against RPE? Is this because they first see the strong cue onset activity in the earlier sessions of cue discrimination task, and then started to observe prolonged activity in the later sessions? Or did this prolonged activity arise first from cue onset and then elongate toward reward, which is in the opposite direction of 'backward shifting'? It is hard to be convinced without a clear rationale. In fact, with sensory uncertainty related arguments as have been made previously in Kim et al. 2020 and Mikhael et al. 2021, I could imagine obtaining these results using RPE encoding. If the authors can simulate a TDRL model and show that this is inconsistent with RPE encoding, that would be more convincing.

We thank the Reviewer for raising this issue. We appreciate the possibility that certain types of RPE model can produce plateau responses, and have added this new section to the Results:

> '**Simple Q-learning RL model cannot account for our observations**
>
> It was clear that the absence laterally of transfer of transient dopamine from outcome to predictive cue and absence medially of a positive outcome response were not in accord with classic RPE models[49]. But this was not so clear for the plateau responses. To assess this issue, we constructed a simple Q-learning model. As expected, this simple model soundly replicated the well-known RPE signals. However, it did not replicate plateau responses, especially the fact that they emerged when the second, unrewarded cue light has added (**Supplementary Fig. 3**). Kim et al.[50] have found that the prolonged ramping dopamine signals reported experimentally[32] can represent RPE signals when temporal discounting ("discount rate" in Ref 2) is a factor. In our data, plateau responses varied in shape from trial to trial (**Supplementary Fig. 3**), but in the aggregate and in most individual mice they were more flattened than ramp-like, and in some animals, even resembled reversed (decreasing) ramps, indicating that temporal discounting was not responsible. Our RL model also did not account for why plateau responses were absent in the cue association task and abruptly emerged when the alternative but never simultaneously presented cue was presented in the sessions (**Supplementary Fig. 3**). As a test for whether the order of task paradigms could be important for these responses, we switched tasks in the Q-learning model so that the cue discrimination preceded rather than followed the cue association sessions. The model did not behave as the mice did: it gave transients at both cue and outcome. More stringently, any RPE model is strongly challenged to explain the decrease of cue-associated phasic response as mice learn the task, or the

> increase of reward-evoked responses in random reward paradigm, even if they incorporate high eligibility trace (lambda) and sensory uncertainty[50,51]. Future studies should address how these issues could be addressed in more complex models that might account for the behavior and dopamine signals reported here.'

We have also added this new paragraph to the Discussion:

> 'It is theoretically possible that a sufficiently complex RL model could be tuned to show a plateau-like component of RPE at intermediate levels of training. Kim et al.[50] showed how temporal discounting can produce upwards ramping RPE responses that resemble ramping dopamine responses that have been reported[10,32,43,44,59-61]. However, such upward ramps were rarely observed in our data. In RL models that endow the agent with a fine sense of the passage of time, such that each time point can be represented as a distinct state, it is also possible to find a small hump in the RPE signal in between the cue and reward delivery that becomes progressively earlier in every trial[1,40]. This hump only occurred in the middle of training, not at the start or end. If the right range of mid-training trials were analyzed, a combination of an upward ramp due to temporal discounting and a moving hump due to progressive transfer of the RPE signal to earlier states in trials could potentially add up to a roughly plateau-like response. We observed plateaus at the end of training, but because we did not keep training our mice after they reached criterion, it is impossible to say based on the present experiments whether the plateaus we observed would persist indefinitely with additional training. Such humps occur at the expense of the reward response, and we did not observe a diminution of the reward response when the plateau components arose. Another potential mechanism that can add a hump to the RPE signal is uncertainty as to exactly when the reward will be delivered[51,62]. In our task, the animal received a very clear reward delivery signal (i.e., the extinguishment of the rewarded cue), so this effect is unlikely to play a role in the present study. Also, a hump due to uncertainty will necessarily be close to the actual reward delivery time, unless the animal systematically overestimates the passage of time between cue and reward. Extensive additional modeling work will be required to determine whether sufficiently complex RL models actually can produce plateau-like RPE signals between cue and reward, and such models might have so many free parameters that they could be fitted to arbitrary data. We thus did not pursue these questions here.'

2. Related to the above point, cue discrimination and single-cue sessions appear to have different reward and cue rates. It's unclear if the emergence of a plateau response has to do with the presence of CS- trials as argued by the authors, or concomitant changes in reward/cue rate in the task.

We apologize for not having stated more clearly that the rates were the same throughout. We ran the same code in cue association and discrimination tasks. In the cue association task, we made the luminescence of non-rewarded cue 0. Thus, reward and cue rates were the same across the two paradigms. We are so sorry that we were not clear. We maintained a consistent number of reward cues and rewards per session across all tasks, except during the probabilistic reward session. As the Reviewer pointed out, we recognized that introducing variations in the presentation of reward cues and rewards could potentially impact underlying motivation. Consequently, it was imperative for us to keep this aspect stable.

> '*Matched reward and cue rates:* All tasks in this study were variants of the same basic discrimination task and were controlled by the same task management code. Reward was given at the same temporal schedule across all tasks which was determined entirely by the task software irrespective of mice's behavior. The only differences among tasks lay in

> the contingencies for delivering or withholding reward and illuminating the cue LEDs. As compared to the common basis, i.e., cue discrimination task, the single-cue task was implemented by simply disabling the non-rewarded cue LED; thus the average rate and time intervals between cue and reward deliveries were also the same. Similarly, the random reward task was implemented by disabling both cue LEDs, the probabilistic reward was implemented by disabling the non-rewarded cue LED and withholding reward on a certain fraction of trials where the nominally rewarded cue was illuminated, and the extinction task was implemented by illuminating the cue LEDs as in the discrimination task but withholding reward on all trials (see lower row of **Fig. 1c**).'

Of course, addressing this would require additional data collection, which may not be easily feasible. So, I'm okay with the authors discussing that the reason for the observed prolonged cue response remains unclear.

We very much appreciate the Reviewer's comments here. We ourselves do not fully understand the plateaus. The impressive work of Kim et al., 2020 makes it evident that RPE can be seen with different forms of RPE-based dopamine release patterns, including the dopamine ramps that we found in Howe et al., 2013. What is observed is that they seem to occur as soon as the mouse must do something more than lick at a light cue trial after trial. In the cue discrimination task, the mice must lick, but there is now an issue, to lick or not, when the non-rewarded cue comes on. The mice after some training suppress this former generalized rule in favor of a specialized rule that states "lick on this, not that trial type". The plateaus do not seem to resemble the dopamine ramps that, for example, we found in locomotor tasks (Howe et al., 2013), in which the ramping dopamine signal seemed to be modulated by proximity to reward, which Kim et al. (2020) then related to RPE signals. We did not use a 3-sec-long cue presentation to allow for detailed analysis of the time course of dopamine concentration between cue and reward, as did Amo et al. (2022) in their impressive paper on the probable existence of "back-propagation" of dopamine signals, so it is difficult to determine whether our plateaus can be explained in similar terms. We are hopeful that further investigation will allow further clarification. We have added this new section to the Results:

> '**Simple Q-learning RL model cannot account for our observations**
>
> It was clear that the absence laterally of transfer of transient dopamine from outcome to predictive cue and absence medially of a positive outcome response were not in accord with classic RPE models[49]. But this was not so clear for the plateau responses. To assess this issue, we constructed a simple Q-learning model. As expected, this simple model soundly replicated the well-known RPE signals. However, it did not replicate plateau responses, especially the fact that they emerged when the second, unrewarded cue light has added (**Supplementary Fig. 3**). Kim et al.[50] have found that the prolonged ramping dopamine signals reported experimentally[32] can represent RPE signals when temporal discounting ("discount rate" in Ref 2) is a factor. In our data, plateau responses varied in shape from trial to trial (**Supplementary Fig. 3**), but in the aggregate and in most individual mice they were more flattened than ramp-like, and in some animals, even resembled reversed (decreasing) ramps, indicating that temporal discounting was not responsible. Our RL model also did not account for why plateau responses were absent in the cue association task and abruptly emerged when the alternative but never simultaneously presented cue was presented in the sessions (**Supplementary Fig. 3**). As a test for whether the order of task paradigms could be important for these responses, we switched tasks in the Q-learning model so that the cue discrimination preceded rather than followed the cue association sessions. The model did not behave as the mice did: it gave transients at both cue and outcome. More stringently, any RPE model is strongly challenged to

> explain the decrease of cue-associated phasic response as mice learn the task, or the increase of reward-evoked responses in random reward paradigm, even if they incorporate high eligibility trace (lambda) and sensory uncertainty[50,51]. Future studies should address how these issues could be addressed in more complex models that might account for the behavior and dopamine signals reported here.'

We have also added this new paragraph to the Discussion:

> ‘It is theoretically possible that a sufficiently complex RL model could be tuned to show a plateau-like component of RPE at intermediate levels of training. Kim et al.[50] showed how temporal discounting can produce upwards ramping RPE responses that resemble ramping dopamine responses that have been reported[10,32,43,44,59-61]. However, such upward ramps were rarely observed in our data. In RL models that endow the agent with a fine sense of the passage of time, such that each time point can be represented as a distinct state, it is also possible to find a small hump in the RPE signal in between the cue and reward delivery that becomes progressively earlier in every trial[1,40]. This hump only occurred in the middle of training, not at the start or end. If the right range of mid-training trials were analyzed, a combination of an upward ramp due to temporal discounting and a moving hump due to progressive transfer of the RPE signal to earlier states in trials could potentially add up to a roughly plateau-like response. We observed plateaus at the end of training, but because we did not keep training our mice after they reached criterion, it is impossible to say based on the present experiments whether the plateaus we observed would persist indefinitely with additional training. Such humps occur at the expense of the reward response, and we did not observe a diminution of the reward response when the plateau components arose. Another potential mechanism that can add a hump to the RPE signal is uncertainty as to exactly when the reward will be delivered[51,62]. In our task, the animal received a very clear reward delivery signal (i.e., the extinguishment of the rewarded cue), so this effect is unlikely to play a role in the present study. Also, a hump due to uncertainty will necessarily be close to the actual reward delivery time, unless the animal systematically overestimates the passage of time between cue and reward. Extensive additional modeling work will be required to determine whether sufficiently complex RL models actually can produce plateau-like RPE signals between cue and reward, and such models might have so many free parameters that they could be fitted to arbitrary data. We thus did not pursue these questions here.’

3. The authors show that the plateau response is not a result of averaging across mice, but it would be useful to also test if it is a result of averaging across trials within individual mice.

We have added a new Figure 3a, showing single trials in a single mouse as an example.

> 'The development of the plateau-like response to the reward-predicting cue was not the product of averaging across mice; it could be seen in individual mice. Nor was it a product of averaging across trials within single sessions. **Fig. 4a** and **b** show the dopamine response averaged across trials for the criterion sessions of cue discrimination and cue reversal for each mouse. **Fig. 3a** illustrates single trial responses from a mouse (animal pa38, which had the fifth largest PC1 amplitude across all mice) in groups of ten trials per plot. The heavy black line in each plot indicates the average of the trials in that same plot. There is considerable volatility in the dopamine signal within each trial. Nonetheless, a relatively consistent pattern of higher maxima and higher minima during the cue, as compared to before the cue, can be seen for every set of ten trials. To get a better view of this pattern, we developed a method for fitting the centerline of the peaks and valleys of a signal by finding the upper envelope determined by the peaks and the lower envelope

> determined by the valleys, and then averaging the upper and lower envelopes. This method is illustrated in **Fig.3b**. First, all 3-point local maxima and minima of the raw dopamine response waveform were found. Then upper and lower envelopes were constructed by, respectively, linear interpolation of the maxima and minima. Finally, the mean of the upper and lower envelopes was calculated, which we refer to as the "midline" of the waveform. **Fig. 3c** shows the same sets of trials as **Fig. 3a**, but showing individual waveform center lines instead of the raw waveforms. The majority of waveform center lines echo the shape of the waveform averaged across trials.'

4. Along similar lines, it is unclear if the observed effects might reflect a strong order effect of the exact experimental conditions that were run. Every animal went through the same sequence of history and thus, it is hard to be sure that the observations here reflect task differences as claimed by the authors, or experimental order effects. Some counterbalancing and controls would be necessary to better interpret the results. Again, since this is a lot of work, I am okay with the authors just discussing this weakness. That said, the methodological details in #2 and in this point diminish my enthusiasm for the strength of the data.

We thank the reviewer tremendously for this comment. We revised the text to discuss the caveats of the sequential training in a fixed order in Results and Methods.

> ‘We are aware of the possible effects of our training regimen with a fixed sequence of paradigms across animals, and that our data might partly reflect such order effects.’
>
> ‘*Possible effects of task order:* Every mouse was trained on the same set of tasks in the same order, and so it is possible that some of the differences reported across tasks might depend on the history of the training rather than on intrinsic differences between the tasks and their corresponding evoked release signaling characteristics. The order of the tasks through reversal discrimination training was chosen partly to minimize the amount of time it took for the mouse to learn each task, and thus to maximize the variety of tasks that we were able to record before the signal quality started to degrade. Testing the effects of task order would have required additional sets of mice beyond the 67 successfully trained here to be trained for each permutation of the task order. We therefore did not attempt to disambiguate this potential confound.’

5. Though the authors contrast the results observed here with Amo et al. 2022, the claim in this paper is not just that there is no systematic backpropagation during the delay period from reward to cue (like Amo et al. 2022) but that the cue response does not increase at all. Since the lack of apparent backpropagation during intermediate timepoints between cue and delay can be explained by a high eligibility trace (lambda), but the current findings cannot be explained by a high lambda, the authors may want to highlight this lack of increasing cue response as a stronger test of reward-to-cue response shift than that considered in prior studies. Related to this, Jeong et al. also claimed to not find the backpropagating bump that Amo et al. found. This may be worth mentioning for completeness. However, this study too was claiming that the cue response increased over learning. Thus, the strength of the observations here is somewhat muddled by the description. I would encourage the authors to highlight this point more.

We yet again thank the Reviewer. We have revised the text to highlight the point that reward and cue response respectively increased and decreased over the course of learning. We have added to the Discussion to explicitly mention his/her point, and we are grateful.

> '**Simple Q-learning RL model cannot account for our observations**

> It was clear that the absence laterally of transfer of transient dopamine from outcome to predictive cue and absence medially of a positive outcome response were not in accord with classic RPE models[49]. But this was not so clear for the plateau responses. To assess this issue, we constructed a simple Q-learning model. As expected, this simple model soundly replicated the well-known RPE signals. However, it did not replicate plateau responses, especially the fact that they emerged when the second, unrewarded cue light has added (**Supplementary Fig. 3**). Kim et al.[50] have found that the prolonged ramping dopamine signals reported experimentally[32] can represent RPE signals when temporal discounting ("discount rate" in Ref 2) is a factor. In our data, plateau responses varied in shape from trial to trial (**Supplementary Fig. 3**), but in the aggregate and in most individual mice they were more flattened than ramp-like, and in some animals, even resembled reversed (decreasing) ramps, indicating that temporal discounting was not responsible. Our RL model also did not account for why plateau responses were absent in the cue association task and abruptly emerged when the alternative but never simultaneously presented cue was presented in the sessions (**Supplementary Fig. 3**). As a test for whether the order of task paradigms could be important for these responses, we switched tasks in the Q-learning model so that the cue discrimination preceded rather than followed the cue association sessions. The model did not behave as the mice did: it gave transients at both cue and outcome. More stringently, any RPE model is strongly challenged to explain the decrease of cue-associated phasic response as mice learn the task, or the increase of reward-evoked responses in random reward paradigm, even if they incorporate high eligibility trace (lambda) and sensory uncertainty[50,51]. Future studies should address how these issues could be addressed in more complex models that might account for the behavior and dopamine signals reported here.'

We have also added a new paragraph in the Discussion:

> ‘It is theoretically possible that a sufficiently complex RL model could be tuned to show a plateau-like component of RPE at intermediate levels of training. Kim et al.[50] showed how temporal discounting can produce upwards ramping RPE responses that resemble ramping dopamine responses that have been reported[10,32,43,44,59-61]. However, such upward ramps were rarely observed in our data. In RL models that endow the agent with a fine sense of the passage of time, such that each time point can be represented as a distinct state, it is also possible to find a small hump in the RPE signal in between the cue and reward delivery that becomes progressively earlier in every trial[1,40]. This hump only occurred in the middle of training, not at the start or end. If the right range of mid-training trials were analyzed, a combination of an upward ramp due to temporal discounting and a moving hump due to progressive transfer of the RPE signal to earlier states in trials could potentially add up to a roughly plateau-like response. We observed plateaus at the end of training, but because we did not keep training our mice after they reached criterion, it is impossible to say based on the present experiments whether the plateaus we observed would persist indefinitely with additional training. Such humps occur at the expense of the reward response, and we did not observe a diminution of the reward response when the plateau components arose. Another potential mechanism that can add a hump to the RPE signal is uncertainty as to exactly when the reward will be delivered[51,62]. In our task, the animal received a very clear reward delivery signal (i.e., the extinguishment of the rewarded cue), so this effect is unlikely to play a role in the present study. Also, a hump due to uncertainty will necessarily be close to the actual reward delivery time, unless the animal systematically overestimates the passage of time between cue and reward. Extensive additional modeling work will be required to determine whether sufficiently complex RL models actually can produce plateau-like RPE signals between cue and

> reward, and such models might have so many free parameters that they could be fitted to arbitrary data. We thus did not pursue these questions here.'

6. One point that I found quite interesting in Jeong et al. 2022 was the increase of dopamine responses to random rewards. They argued that this is inconsistent with RPE since the initial presentations of rewards are the most unpredicted. The authors similarly find here that random reward responses increase across training. This appears to also be in violation of TD-RPE. The authors may want to highlight this finding. Based on my understanding, an RPE (received minus predicted reward) signal can increase only if prediction is negative or received reward changes. Neither appears to be the case here, suggesting that this is an interesting result to highlight.

We yet again thank the Reviewer. We have revised the text to highlight the point. The new initial paragraph for the Discussion includes this text:

> 'Also, in both centromedial and centrolateral striatum, the dopamine response to random reward increased with training, whereas an RPE signal would decrease with training. The RPE model is thus not sufficient to account for these data.'
>
> '**Simple Q-learning RL model cannot account for our observations**
> It was clear that the absence laterally of transfer of transient dopamine from outcome to predictive cue and absence medially of a positive outcome response were not in accord with classic RPE models[49]. But this was not so clear for the plateau responses. To assess this issue, we constructed a simple Q-learning model. As expected, this simple model soundly replicated the well-known RPE signals. However, it did not replicate plateau responses, especially the fact that they emerged when the second, unrewarded cue light has added (**Supplementary Fig. 3**). Kim et al.[50] have found that the prolonged ramping dopamine signals reported experimentally[32] can represent RPE signals when temporal discounting ("discount rate" in Ref 2) is a factor. In our data, plateau responses varied in shape from trial to trial (**Supplementary Fig. 3**), but in the aggregate and in most individual mice they were more flattened than ramp-like, and in some animals, even resembled reversed (decreasing) ramps, indicating that temporal discounting was not responsible. Our RL model also did not account for why plateau responses were absent in the cue association task and abruptly emerged when the alternative but never simultaneously presented cue was presented in the sessions (**Supplementary Fig. 3**). As a test for whether the order of task paradigms could be important for these responses, we switched tasks in the Q-learning model so that the cue discrimination preceded rather than followed the cue association sessions. The model did not behave as the mice did: it gave transients at both cue and outcome. More stringently, any RPE model is strongly challenged to explain the decrease of cue-associated phasic response as mice learn the task, or the increase of reward-evoked responses in random reward paradigm, even if they incorporate high eligibility trace (lambda) and sensory uncertainty[50,51]. Future studies should address how these issues could be addressed in more complex models that might account for the behavior and dopamine signals reported here.'

We have also added this new paragraph to the Discussion:

> 'It is theoretically possible that a sufficiently complex RL model could be tuned to show a plateau-like component of RPE at intermediate levels of training. Kim et al.[50] showed how temporal discounting can produce upwards ramping RPE responses that resemble ramping dopamine responses that have been reported[10,32,43,44,59-61]. However, such upward ramps were rarely observed in our data. In RL models that endow the agent with a fine sense of the passage of time, such that each time point can be represented as a

> distinct state, it is also possible to find a small hump in the RPE signal in between the cue and reward delivery that becomes progressively earlier in every trial[1,40]. This hump only occurred in the middle of training, not at the start or end. If the right range of mid-training trials were analyzed, a combination of an upward ramp due to temporal discounting and a moving hump due to progressive transfer of the RPE signal to earlier states in trials could potentially add up to a roughly plateau-like response. We observed plateaus at the end of training, but because we did not keep training our mice after they reached criterion, it is impossible to say based on the present experiments whether the plateaus we observed would persist indefinitely with additional training. Such humps occur at the expense of the reward response, and we did not observe a diminution of the reward response when the plateau components arose. Another potential mechanism that can add a hump to the RPE signal is uncertainty as to exactly when the reward will be delivered[51,62]. In our task, the animal received a very clear reward delivery signal (i.e., the extinguishment of the rewarded cue), so this effect is unlikely to play a role in the present study. Also, a hump due to uncertainty will necessarily be close to the actual reward delivery time, unless the animal systematically overestimates the passage of time between cue and reward. Extensive additional modeling work will be required to determine whether sufficiently complex RL models actually can produce plateau-like RPE signals between cue and reward, and such models might have so many free parameters that they could be fitted to arbitrary data. We thus did not pursue these questions here.’

7. To be honest, I felt that it’s a missed opportunity that the authors did not discuss (i.e., Discussion section) the difference between medial vs. lateral striatum.

We would like to thank the Reviewer for helping us with this valuable suggestion. We have now revised the text to further discuss the difference between medial and lateral striatum.

> 'Dopamine waves have been reported to travel in a lateral to medial direction when mice learn Pavlovian tasks rather than instrumental ones[19]. Also, lateral and medial parts of the dorsal striatum have been found to receive projections from different molecular-subtypes of dopaminergic neurons; the calbindin-positive type signals RPE, whereas the Anxa1-positive type encodes the acceleration of locomotion[65]. Thus, the differential dopamine dynamics observed in this study could be attributed to the directional dopamine wave or the differential contribution of dopaminergic cell subtypes to the ambient dopamine content of the different sub-regions. Delineating the functions of these neuromodulatory and neurochemical gradients awaits future study.'

I think these data has great potential to discuss this matter, instead of just focusing on RPE. On a related note, I thought Figure 3e is very interesting. Even in non-learners, there was a clear distinction in dopamine release to cues in DLS, but not in DMS. Is this related to the motor control aspect of DMS?

We are sorry, but we do not know for sure, and therefore did not mention this alternative. But we thank the Reviewer—this is indeed interesting. We now mention the possibility that the difference in response could be related to some aspect of motor control:

> ‘We cannot account for the mediolateral differences in the non-learners (**Fig. 4e**); one possibility is that this difference is related to motor learning.’

8. Related to the above finding, it is already known that DLS has higher DAT expression and hence, faster dynamics than DMS. Can this instruct the conclusion in Figure 5 (e.g., DMS can appear to have more plateau-like activity because of its slow timescale)?

This is definitely one very interesting possibility. Of course, there are many others, and therefore we held off in speculating. But we now have added:

> ‘One possibility to raise here is that the especially emphasized plateau in the medial sites is related to the lower concentration of dopamine transporter (uptake) than in the lateral sites, and so the medial sites have longer (slower) plateaus.’

Minor

1. In the discussion, the authors describe the Jeong et al. study and the current study as measuring from dopamine axons. Both studies use dopamine release measurements and not dopamine axon GCaMP dynamics. It may be better to be more careful about this description.

We thank the Reviewer for raising the point. We have revised the text to be clear and accurate.

> ‘For dopamine release recorded in the ventral striatum/nucleus accumbens, Jeong et al.[37], with a series of conditioning tasks, have found inconsistencies between dopamine release signals there and the predictions of RPE formulations. These favor what the authors term as a retrospective causal learning algorithm. The recordings by Jeong et al., like our recordings, were made with the aid of dopamine sensors, not with microelectrodes recording the spike activity of dopamine-containing cell bodies as in the original studies linking dopamine dynamics to RPE. Such discrepancies could be accounted for by findings of the Uchida and Watabe-Uchida and Uchida groups (e.g., Ref 40; see also Ref 39).’

2. The authors state that “The recordings made by Jeong et al., like our recordings, were made from dopamine-containing axons, not from dopamine cell bodies as in the original studies linking dopamine dynamics to RPE. These discrepancies could be accounted for according the Uchida and Watabe-Uchida and Uchida groups (e.g., Ref 40; see also Ref 39).” This sentence seems a bit unclear to me. Are the authors stating that Ref 40 explains the discrepancies between RPE and dopamine release data from Jeong et al., or that Ref 40 found no discrepancy between cell body activity and dopamine axon/release dynamics? I think they might potentially mean the latter, but the sentence reads like the former. Since the Jeong et al. paper is more recent than Ref 40, I am confused about how Ref 40 accounts for the Jeong et al. results. Perhaps I am misreading this sentence but it is good to be clear.

We are so sorry for the confusion. This was very poorly worded; we now realize, thanks to the Reviewer. What we meant to say was: These issues are general to the field as deeper insights are gained about the properties and functions of dopamine-containing neurons and their effects at their targets. These issues have recently been discussed by the Watabe-Uchida and Uchida groups in publications both before and after the Jeong et al. paper was published.

> 'Such discrepancies could be accounted for by findings of the Uchida and Watabe-Uchida and Uchida groups (e.g., Ref 40; see also Ref 39).'

Reviewer #3 (Remarks today the Author):

Several recent studies have challenged the classical hypothesis that midbrain dopamine (DA) neurons signal reward prediction error (RPE) to the striatum. The authors tested this issue for the dorsal striatum by training mice consecutively on a series of visual cue-outcome conditioning tasks and recording dopamine release responses with chemosensor probes throughout the time that the mice were learning the tasks. The Introduction beautifully summarizes the rationale of the aims. Experiments are in most parts very well thought out.

Recording sites are mostly within a small anterior dorsal striatal subregion.
The results clearly show that DA release profiles in the dorsal striatum undergo very dynamic learning-related changes. An amazing finding is that within the small region DA signals show clear location dependent differences which are consistent among animals. These different release profiles were also selective for different versions of associative cue-outcome conditioning. Importantly, they found no evidence for a transition from outcome to cue signaling that a hallmark of temporal difference reinforcement learning as applied to the substantia nigra pars compacta (SNpc) activity reported in previous studies. Results greatly extend recent concept that SNpc DA neurons and the striatum have heterogeneous and specialized responses to task variables during complex behavior. I totally agree with the authors suggestion that "further refinement and extension of reinforcement learning algorithms is needed to account for spatiotemporal DA release dynamics in the dorsal striatum".

The results are noteworthy and significant for the advancement of the basal ganglia research. Comments below are to clarify some minor issues in the manuscript.

The limitations of the studies are clearly described. One such limitation is that possible stimulation of DA fibers by acetylcholine in the striatum in vivo is speculative. Two very recent Nature articles (s41586-023-06492-9_reference and s41586-023-05995-9) describing acetylcholine and DA dynamics in the striatum are relevant to this issue.

We thank the Reviewer for this comment. We have revised the Introduction and Discussion sections to refer the previous results by Krok et al., 2023 and Chantranupong et al., 2023.

> Introduction: 'The release can exhibit low frequency oscillations during or even without task engagement, gated by extrinsic striatal afferents[20].'
>
> Discussion: 'Further dynamics of intraneuronal networks in the striatum, such as the activation of dopaminergic fibers by acetylcholine released from cholinergic interneuron[20,58], surely must contribute.'

Reason to target two sites within the small dorsal striatum become clear in later part of the manuscript. It could be mentioned in the Method or in Introduction why the medial site (AP: +1.0 mm, ML: +1.5 mm, DV: −2.7 mm from bregma) and lateral site (AP: +1.0 mm, ML: +1.9 mm, DV: −3.0 mm from bregma) which have only 0.5mm separation were targeted.

We thank the Reviewer for this comment. We fully agree that this heterogeneity across even a small more medial and more lateral part of the striatum with limited anteroposterior difference is remarkable. This speaks to the profound need for spatially dense and extensive recordings such as now being done by the Howe laboratory. We look forward to their findings.

We made this clear in the Methods section.

> 'It should be noted that there was considerable scatter in the final positions of the probe tips (see **Figs. 1f** and **6c, d**).'

Also, please describe the diameter and the shape of optic probe tip.

We thank the Reviewer. We now have described in the Methods section.

> ‘via an optic fiber of circular cross-section having a 200-μm diameter‘

Most of recording sites shown in F5 and supplemental 1 are in the medial sector of the striatum, if “Recording sites were classified as "medial" if the distance from the midline to the tip of the probe was less than 0.6 of the distance from the midline to the lateral edge of the striatum.”

The value of 0.6 associated with the ML location does not represent a distance from the midline, but rather a ratio scale that we utilized for identifying relative locations. This ratio was derived by comparing the distance from the midline to the outermost lateral part of the striatum with the distance from the midline to the recording tip. We used this ratio to segregate our data into relatively medial and lateral positions based on a ratio value of 0.6. In actual anatomical terms, our lateral area is predominantly concentrated in the dorsocentral region, rather than situated in the far lateral site. We have added the following material to the Methods to clarify this point.

> ‘We defined four reference points in each coronal section: the most medial point of striatum (x1, y1; usually about 1 mm from midline), the most lateral point of striatum (x2, y2), the most dorsal point of striatum (x3, y3), and the most dorsal point of anterior commissure (x4, y4). Designating the tip of probe as (x5, y5), the standardized medial-lateral coordinate of the probe tip was calculated as (x5 – x1) / (x2 – x1), and the standardized depth coordinate of the probe tip was calculated as (y5 – y3) / (y4 – y3). We refer to this coordinate system as "relative position" or "standardized coordinate" (shown, e.g., in **Fig. 6 c** and **d**).
>
> Recording sites were classified as "medial" if the distance from the midline to the tip of the probe was less than 0.6 of the distance from the midline to the lateral edge of the striatum. Note that this measurement was not done according to the "standardized coordinate" system. Depending on the A-P plane of section, the division between "medial" and "lateral" corresponded to around 0.4 to 0.5 M-L in the "standardized coordinate" system. A-P coordinates were determined by comparing histological sections to the atlas[69] and are given relative to bregma as in the atlas. In **Fig. 1b** and **f**, M-L coordinates are our standardized coordinate ratio, and D-V coordinates are the distance (mm) from dorsal surface of the striatum as it appears in same section containing the track.’

Fig. 1b, please give the size of scale bar.

Thank you! We have added this text to the figure legend:

> 'Scale bar: 500 μm.'

Fig. 1b and 1F, DMs are in mm. Please describe the location of the reference point (DV=0.0).

We have modified the text as follows:

> ‘A-P coordinates were determined by comparing histological sections to the atlas[69] and are given relative to bregma as in the atlas. In **Fig. 1b** and **f**, M-L coordinates are our standardized coordinate ratio, and D-V coordinates are the distance (mm) from dorsal surface of the striatum as it appears in same section containing the track.’

Fig. 1f, Mixed use of mm and standardized coordinates for probe locations is confusing. Methods to standardize coordinates for ML and DV locations are described but not for AP used in Fig. 5 and supplemental 1.

We apologize for, and deeply regret, this confusion, and for the omission regarding AP coordinates. The histological analyses were done by several people without sufficient coordination. We have made the following changes to clarify.

> 'We defined four reference points in each coronal section: the most medial point of striatum (x1, y1; usually about 1 mm from midline), the most lateral point of striatum (x2, y2), the most dorsal point of striatum (x3, y3), and the most dorsal point of anterior commissure (x4, y4). Designating the tip of probe as (x5, y5), the standardized medial-lateral coordinate of the probe tip was calculated as (x5 – x1) / (x2 – x1), and the standardized depth coordinate of the probe tip was calculated as (y5 – y3) / (y4 – y3). We refer to this coordinate system as "relative position" or "standardized coordinate" (shown, e.g., in **Fig. 6 c** and **d**).
>
> Recording sites were classified as "medial" if the distance from the midline to the tip of the probe was less than 0.6 of the distance from the midline to the lateral edge of the striatum. Note that this measurement was not done according to the "standardized coordinate" system. Depending on the A-P plane of section, the division between "medial" and "lateral" corresponded to around 0.4 to 0.5 M-L in the "standardized coordinate" system. A-P coordinates were determined by comparing histological sections to the atlas[69] and are given relative to bregma as in the atlas. In **Fig. 1b** and **f**, M-L coordinates are our standardized coordinate ratio, and D-V coordinates are the distance (mm) from dorsal surface of the striatum as it appears in same section containing the track.'

Reviewer #4 (Remarks to the Author):

Kim et al. recorded dopamine sensor signals with a single fiber in the dorsomedial striatum (DMS) or dorsolateral striatum (DLS) in head-fixed mice. The mice experienced a series of tasks with water reward, some with visual cues, a light on the left or right. Authors found great individual variability in the sensor signals, some explained by fiber locations, sexes and learning extents. From these complicated results, authors concluded that dopamine release in the dorsal striatum is not consistent with reward prediction error.

Since signature studies in monkeys by Montagues et al., 1996 and Shultz et al., 1997, dopamine neuron activity had often been interpreted as reward prediction errors. However, many recent rodent studies as well as some monkey studies discovered that dopamine activity is more complicated than it had been believed, and that dopamine neurons are diverse at molecular levels as well as at functional levels, some of which correspond to diversity of projection targets. Even though the precise interpretation is different among studies, there is a consensus now that dopamine neurons are not uniform. This study by Kim et al. adds another example to yield such a conclusion. Although the presented results are potentially interesting, most findings are overlapped with previous studies.

We thank the Reviewer for pointing out this issue. Yes, there are overlaps, but we showed for the first time, as far as we know, that the striatal dopamine release can strictly deviate from the RPE model; cue-evoked response decreases and reward-evoked response increases as mice learn. We made this point clear in the Results and Discussion sections. We are aware of the 'complexity', but we, as fellow scientists, have the conviction that principles will in time emerge to give simplifying and deep accounting of these within a rational framework.

> '**Simple Q-learning RL model cannot account for our observations**
>
> It was clear that the absence laterally of transfer of transient dopamine from outcome to predictive cue and absence medially of a positive outcome response were not in accord with classic RPE models[49]. But this was not so clear for the plateau responses. To assess this issue, we constructed a simple Q-learning model. As expected, this simple model soundly replicated the well-known RPE signals. However, it did not replicate plateau responses, especially the fact that they emerged when the second, unrewarded cue light has added (**Supplementary Fig. 3**). Kim et al.[50] have found that the prolonged ramping dopamine signals reported experimentally[32] can represent RPE signals when temporal discounting ("discount rate" in Ref 2) is a factor. In our data, plateau responses varied in shape from trial to trial (**Supplementary Fig. 3**), but in the aggregate and in most individual mice they were more flattened than ramp-like, and in some animals, even resembled reversed (decreasing) ramps, indicating that temporal discounting was not responsible. Our RL model also did not account for why plateau responses were absent in the cue association task and abruptly emerged when the alternative but never simultaneously presented cue was presented in the sessions (**Supplementary Fig. 3**). As a test for whether the order of task paradigms could be important for these responses, we switched tasks in the Q-learning model so that the cue discrimination preceded rather than followed the cue association sessions. The model did not behave as the mice did: it gave transients at both cue and outcome. More stringently, any RPE model is strongly challenged to explain the decrease of cue-associated phasic response as mice learn the task, or the increase of reward-evoked responses in random reward paradigm, even if they incorporate high eligibility trace (lambda) and sensory uncertainty[50,51]. Future studies should address how these issues could be addressed in more complex models that might account for the behavior and dopamine signals reported here.'

Most importantly, main findings should be verified with proper controls and analyses. There are additional concerns about data quality and data exclusion.

Major concerns (not exhaustive) We are very sorry for to have led to this highly negative style of comments in the Reviewer's Subheading Reviewer's comments. We do our best below to reply to his/her comments., many of which were extremely helpful and have led us to many changes in the manuscript.

1. After several different tasks, dopamine release both in DMS and DLS became consistent with RPE. This was interpreted by authors that the later tasks are more cognitively demanding. Although this is an interesting idea, it is similarly possible that animals just needed to be trained longer in each task for dopamine neurons to exhibit RPE-like activity patterns.

We appreciate the Reviewer's comment pointing out one of the core issues of this study. It is indeed important to know what factors contribute to or are required for the emergence of an RPE-like pattern. Overtraining might be one of these; in other words, the RPE model might be able to explain the later stage of learning but not the earlier stage of learning.

Thus, our main point would still hold, i.e., dopamine signaling can be strikingly non-RPE-like. The two RPE-like patterns, i.e., the response amplitude scaled by the reward probability and the more medial region's reward response early in cue reversal training, were unique in the present data set. Even these RPE-like responses still show no sign of the "transfer of activation" described by

Schultz (1998). We revised the main text to explain the bases of our claims. We again emphasize that the pioneering findings of Schultz and his colleagues referred to electrophysiological recording in the cell-body region of the SNpc, whereas the striatal recordings reported here were obtained by the chemogenetic dopamine sensors developed by Patriarchi et al., 2018 and Sun et al., 2018. We have added text to the Discussion to highlight this issue.

> 'In rare occasions in our dataset, the dopamine responses exhibited RPE-like patterns, i.e., responses to the probabilistic reward and reward response early in cue reversal training recorded from centromedial striatum. The factors to shape dopamine response to be RPE-like are unknown, and possibly include cognitive demands or effects of overtraining. We await future studies to identify these factors.'

Multiple studies show that simple behavioral changes such as anticipatory licking are observed even before visible changes in population activity of dopamine neurons (ref. Menegas et al., 2017). Authors should have controls to train longer without switching tasks to determine whether cognitive demanding affects RPE-like signals in dopamine.

The Reviewer raises a central issue that we have thought about as the findings came in and were not all in accord with standard RPE models. But models are models that are not meant to be permanent. Their special value is as guides to go deeper into the findings. This argument presents an intriguing perspective to be explored in the future study. In this study, our main point is that 'dopamine response pattern can be non-RPE-like', rather than 'cognitive demand or overtraining make RPE-like response manifested'. It is possible that our hypothesis is not correct, namely that cognitive demand (or over-training) is a leading candidate factor to make the dopamine response RPE-like. However, our claims still stand. If RPE is the model to explain our dataset, the model should be able to explain the data in early training as well as later training stages. This study presents the counterexample. The factors to make RPE-like dopamine pattern manifested will be explored in future studies. Overtraining may represent a form of learning that is distinct from initial task learning and has been shown to rely on devaluation-resistant habit formation rather than devaluation-sensitive goal-oriented learning. The present study focuses on initial learning, not habit formation. We revised the main text to explain the bases of our claims. In summary, we again thank the Reviewer for his/her penetrating comment.

> 'In rare occasions in our dataset, the dopamine responses exhibited RPE-like patterns, i.e., responses to the probabilistic reward and reward response early in cue reversal training recorded from centromedial striatum. The factors to shape dopamine response to be RPE-like are unknown, and possibly include cognitive demands or effects of overtraining. We await future studies to identify these factors.'
>
> '**Simple Q-learning RL model cannot account for our observations**
> It was clear that the absence laterally of transfer of transient dopamine from outcome to predictive cue and absence medially of a positive outcome response were not in accord with classic RPE models[49]. But this was not so clear for the plateau responses. To assess this issue, we constructed a simple Q-learning model. As expected, this simple model soundly replicated the well-known RPE signals. However, it did not replicate plateau responses, especially the fact that they emerged when the second, unrewarded cue light has added (**Supplementary Fig. 3**). Kim et al.[50] have found that the prolonged ramping dopamine signals reported experimentally[32] can represent RPE signals when temporal discounting ("discount rate" in Ref 2) is a factor. In our data, plateau responses varied in shape from trial to trial (**Supplementary Fig. 3**), but in the aggregate and in most individual mice they were more flattened than ramp-like, and in some animals, even resembled reversed (decreasing) ramps, indicating that temporal discounting was not responsible.

> Our RL model also did not account for why plateau responses were absent in the cue association task and abruptly emerged when the alternative but never simultaneously presented cue was presented in the sessions (**Supplementary Fig. 3**). As a test for whether the order of task paradigms could be important for these responses, we switched tasks in the Q-learning model so that the cue discrimination preceded rather than followed the cue association sessions. The model did not behave as the mice did: it gave transients at both cue and outcome. More stringently, any RPE model is strongly challenged to explain the decrease of cue-associated phasic response as mice learn the task, or the increase of reward-evoked responses in random reward paradigm, even if they incorporate high eligibility trace (lambda) and sensory uncertainty[50,51]. Future studies should address how these issues could be addressed in more complex models that might account for the behavior and dopamine signals reported here.'

We have also added this new paragraph to the Discussion:

> 'It is theoretically possible that a sufficiently complex RL model could be tuned to show a plateau-like component of RPE at intermediate levels of training. Kim et al.[50] showed how temporal discounting can produce upwards ramping RPE responses that resemble ramping dopamine responses that have been reported[10,32,43,44,59-61]. However, such upward ramps were rarely observed in our data. In RL models that endow the agent with a fine sense of the passage of time, such that each time point can be represented as a distinct state, it is also possible to find a small hump in the RPE signal in between the cue and reward delivery that becomes progressively earlier in every trial[1,40]. This hump only occurred in the middle of training, not at the start or end. If the right range of mid-training trials were analyzed, a combination of an upward ramp due to temporal discounting and a moving hump due to progressive transfer of the RPE signal to earlier states in trials could potentially add up to a roughly plateau-like response. We observed plateaus at the end of training, but because we did not keep training our mice after they reached criterion, it is impossible to say based on the present experiments whether the plateaus we observed would persist indefinitely with additional training. Such humps occur at the expense of the reward response, and we did not observe a diminution of the reward response when the plateau components arose. Another potential mechanism that can add a hump to the RPE signal is uncertainty as to exactly when the reward will be delivered[51,62]. In our task, the animal received a very clear reward delivery signal (i.e., the extinguishment of the rewarded cue), so this effect is unlikely to play a role in the present study. Also, a hump due to uncertainty will necessarily be close to the actual reward delivery time, unless the animal systematically overestimates the passage of time between cue and reward. Extensive additional modeling work will be required to determine whether sufficiently complex RL models actually can produce plateau-like RPE signals between cue and reward, and such models might have so many free parameters that they could be fitted to arbitrary data. We thus did not pursue these questions here.'

2. Single animals were used to test dopamine activity in multiple tasks without considering any effects of history. For example, mice were intensively trained with Operant tasks (lick to trigger water delivery) with almost fixed inter-trial-interval (ITI 6-8s) before starting random reward sessions.

We thank the Reviewer for the suggestions. We have revised the manuscript to discuss the possible effects of training order history and made it clear that the lick activity-dependent (LAD) reward habituation (Operant task) was run typically only 1-2 days for each mouse. We apologize for not having emphasized that this was very brief, needed only to habituate the mice to the task context and environment. We have now tried to clarify this problem in the following and again thank the Reviewer:

> Results: 'We are aware of the possible effects of our training regimen with a fixed sequence of paradigms across animals, and that our data might partly reflect such order effects.'
>
> Methods: '*Possible effects of task order:* Every mouse was trained on the same set of tasks in the same order, and so it is possible that some of the differences reported across tasks might depend on the history of the training rather than on intrinsic differences between the tasks and their corresponding evoked release signaling characteristics. The order of the tasks through reversal discrimination training was chosen partly to minimize the amount of time it took for the mouse to learn each task, and thus to maximize the variety of tasks that we were able to record before the signal quality started to degrade. Testing the effects of task order would have required additional sets of mice beyond the 67 successfully trained here to be trained for each permutation of the task order. We therefore did not attempt to disambiguate this potential confound.'
>
> 'The 1-hr LAD habituation (Operant) sessions continued daily until mice actively consumed more than 150 droplets per session. Most of the mice required 1-2 habituation sessions, but some mice required more.'

This is a quite complicated situation for animals; they are trained to expect water every time they lick around 6-8s after the previous water, and then now such a lick is not valid anymore, but instead they have to wait longer (ITI 8-48s) for passive delivery. Indeed, the lick trace in Figure 1d columns 1&2 show that what the animal learned was to lick less during ITI. While authors indicate that their observation in dopamine activity is not consistent with TD error, it is not clear what kind of TD learning should explain this learning and what kind of TD errors are expected to explain dopamine activity in such a complicated situation, switching between Operant and passive behaviors.

We want to especially thank the Reviewer for this comment. This point, i.e., discrepancy from TD error signal, was also raised by other Reviewers. We have revised the text to clarify the point.

> '**Simple Q-learning RL model cannot account for our observations**
>
> It was clear that the absence laterally of transfer of transient dopamine from outcome to predictive cue and absence medially of a positive outcome response were not in accord with classic RPE models[49]. But this was not so clear for the plateau responses. To assess this issue, we constructed a simple Q-learning model. As expected, this simple model soundly replicated the well-known RPE signals. However, it did not replicate plateau responses, especially the fact that they emerged when the second, unrewarded cue light has added (**Supplementary Fig. 3**). Kim et al.[50] have found that the prolonged ramping dopamine signals reported experimentally[32] can represent RPE signals when temporal discounting ("discount rate" in Ref 2) is a factor. In our data, plateau responses varied in shape from trial to trial (**Supplementary Fig. 3**), but in the aggregate and in most individual mice they were more flattened than ramp-like, and in some animals, even resembled reversed (decreasing) ramps, indicating that temporal discounting was not responsible. Our RL model also did not account for why plateau responses were absent in the cue association task and abruptly emerged when the alternative but never simultaneously presented cue was presented in the sessions (**Supplementary Fig. 3**). As a test for whether the order of task paradigms could be important for these responses, we switched

tasks in the Q-learning model so that the cue discrimination preceded rather than followed the cue association sessions. The model did not behave as the mice did: it gave transients at both cue and outcome. More stringently, any RPE model is strongly challenged to explain the decrease of cue-associated phasic response as mice learn the task, or the increase of reward-evoked responses in random reward paradigm, even if they incorporate high eligibility trace (lambda) and sensory uncertainty[50,51]. Future studies should address how these issues could be addressed in more complex models that might account for the behavior and dopamine signals reported here.'

Similarly, those animals were then trained to associate a single cue and water, after intensive random water sessions. Again, under the history of high expectation of random water, introduction of a cue in this task trains animals to hold licking during ITI, rather than to make them to lick more after a cue (Figure 1d column 3&4). What is authors' prediction for TD errors in this task?

As the Reviewer kindly points out, this is indeed a complicated learning situation, not only for animals but also for RL models. We have added new Supplementary Figure 3 specifically comparing the RPE results from our simple Q-learning model with our dopamine photometry results, and accompanying new text in the Results section.

> **'Supplementary Fig. 3**. **Comparison of model RPE results with experimental results for dopamine release responses.** See Supplementary Text. **a** State transition diagram for the Q-learning model. "S1" - "S5": model states. Each arrow exiting from each state is annotated with the action represented and its net reward value. **b-d** First column shows trial-averaged RPE values from the Q-learning model; second and third columns are reproduced from main **Fig. 2a** for convenience of comparison. **b** Random reward responses. The model was trained only on the random reward task. Model trials were aligned on reward delivery (time 0) and averaged for comparison to the trial-averaged dopamine responses. First licks normally occurred on the next time step. Blue: first 3 trials; red: last 10 trials. **c** Cue discrimination training. Training began directly with the cue discrimination task. Model trials were aligned on cue onset (time 0). Cue termination and reward delivery were 1.5 sec after cue onset. Blue: first 10 trials; red: last 10 trials. Note that the discrimination task plots copied from **Fig. 2a** show cue onset at time 0.5. **d** Reversal discrimination training, aligned as in b. Reversal training followed cue discrimination training in order to show the reversal dynamics. Blue: first 10 trials; red: last 10 trials.'
>
> '**Simple Q-learning RL model cannot account for our observations**
> It was clear that the absence laterally of transfer of transient dopamine from outcome to predictive cue and absence medially of a positive outcome response were not in accord with classic RPE models[49]. But this was not so clear for the plateau responses. To assess this issue, we constructed a simple Q-learning model. As expected, this simple model soundly replicated the well-known RPE signals. However, it did not replicate plateau responses, especially the fact that they emerged when the second, unrewarded cue light has added (**Supplementary Fig. 3**). Kim et al.[50] have found that the prolonged ramping dopamine signals reported experimentally[32] can represent RPE signals when temporal discounting ("discount rate" in Ref 2) is a factor. In our data, plateau responses varied in shape from trial to trial (**Supplementary Fig. 3**), but in the aggregate and in most individual mice they were more flattened than ramp-like, and in some animals, even resembled reversed (decreasing) ramps, indicating that temporal discounting was not responsible. Our RL model also did not account for why plateau responses were absent in the cue association task and abruptly emerged when the alternative but never simultaneously

> presented cue was presented in the sessions (**Supplementary Fig. 3**). As a test for whether the order of task paradigms could be important for these responses, we switched tasks in the Q-learning model so that the cue discrimination preceded rather than followed the cue association sessions. The model did not behave as the mice did: it gave transients at both cue and outcome. More stringently, any RPE model is strongly challenged to explain the decrease of cue-associated phasic response as mice learn the task, or the increase of reward-evoked responses in random reward paradigm, even if they incorporate high eligibility trace (lambda) and sensory uncertainty[50,51]. Future studies should address how these issues could be addressed in more complex models that might account for the behavior and dopamine signals reported here.'

In the next task, another cue for no outcome was introduced, after animals already learned a reward-predicting cue. Why, then, should water responses in dopamine in this task be moved from outcome to a reward-predicting cue again as authors predicted from TD errors?

We apologize once more for being so unforthcoming regarding our expectations for RPE signals. Our expectation was that there would be little response to the cue when it was first introduced at the start of cue conditioning training, and large response to the outcome. In the cue discrimination task, after mice had already learned the association between the cue and outcome, there should not be a transfer of activation for the CS because that should have already happened during cue conditioning (which we did not in fact observe). For the alternative cue never paired with reward, we need to make an assumption as to whether the mouse generalizes from one cue to the other cue, and this might well vary from mouse to mouse. If the mouse does generalize, then there should be observed a transfer of **in**activation, i.e., firstly there is a negative outcome response when reward is not delivered, and then gradually this suppression of signals should be transferred to the presentation of the cue not predictive of reward. If the mouse does not generalize, then the cue should initially produce either no response or a negative response, and there should be no response to the absence of reward. However, the observed results were dramatically different from any of those expected patterns. We have made several changes to the text to highlight this discrepancy.

> 'We did see a positive response to cue onset at the very beginning of discrimination training, and medially an absence of response at reward delivery, both of which would be expected from the RPE model as a result of cue conditioning training; but we note that these did not evolve during cue conditioning as expected for RPE. The evolution of these responses, with discrimination training evoking the same or nearly the same responses but with a plateau phase added, cannot readily be accounted for in terms of RPE.'
>
> 'Such humps occur at the expense of the reward response, and we did not observe a diminution of the reward response when the plateau components arose.'

The example animal in Figure 1d column 5&6 even learned to suppress licking after a reward-predicting cue, instead of increasing anticipatory licking.

This is a most acute and intriguing observation, for which we thank the Reviewer. We are only beginning to try to make sense of the entire phenomenon of anticipatory licking, which is rather difficult to understand, especially the question of when and why anticipatory licking is initiated, and when it is suppressed. We hope to make progress on this issue in future work. We have added these remarks to the Discussion:

> 'Another phenomenon that is difficult to explain in terms of classic reinforcement learning theory is anticipatory licking. In any state in which reward is not available, licking produces a net loss, and so the model will learn to wait instead of to lick in those states. Thus, anticipatory licking seems to go outside the basic reinforcement learning framework. Additional innovative modeling work will be required to find an appropriate way to deal with it.'

Finally, authors observed RPE-like dopamine signals in later tasks in reversal

Again, we thank the Reviewer to point this out. Yes, it is true that Fig. 2a does show one feature that is reminiscent of classic RPE, namely the large positive change in dopamine at the start of reversal training in response to reward delivery following the newly rewarded cue, which was the unrewarded cue in the previous sessions. However, the other features of dopamine response during reversal training diverged from the expectations of RPE in similar ways to previous blocks of training. We have added the following paragraph discussing the RPE-type response seen in this study. We have made the following changes to the text to highlight the specific aspects of dopamine response to particular phases of task training that diverged from expectations base on the RPE interpretation:

> '**Simple Q-learning RL model cannot account for our observations**
>
> It was clear that the absence laterally of transfer of transient dopamine from outcome to predictive cue and absence medially of a positive outcome response were not in accord with classic RPE models[49]. But this was not so clear for the plateau responses. To assess this issue, we constructed a simple Q-learning model. As expected, this simple model soundly replicated the well-known RPE signals. However, it did not replicate plateau responses, especially the fact that they emerged when the second, unrewarded cue light has added (**Supplementary Fig. 3**). Kim et al.[50] have found that the prolonged ramping dopamine signals reported experimentally[32] can represent RPE signals when temporal discounting ("discount rate" in Ref 2) is a factor. In our data, plateau responses varied in shape from trial to trial (**Supplementary Fig. 3**), but in the aggregate and in most individual mice they were more flattened than ramp-like, and in some animals, even resembled reversed (decreasing) ramps, indicating that temporal discounting was not responsible. Our RL model also did not account for why plateau responses were absent in the cue association task and abruptly emerged when the alternative but never simultaneously presented cue was presented in the sessions (**Supplementary Fig. 3**). As a test for whether the order of task paradigms could be important for these responses, we switched tasks in the Q-learning model so that the cue discrimination preceded rather than followed the cue association sessions. The model did not behave as the mice did: it gave transients at both cue and outcome. More stringently, any RPE model is strongly challenged to explain the decrease of cue-associated phasic response as mice learn the task, or the increase of reward-evoked responses in random reward paradigm, even if they incorporate high eligibility trace (lambda) and sensory uncertainty[50,51]. Future studies should address how these issues could be addressed in more complex models that might account for the behavior and dopamine signals reported here.'

We have also added this new paragraph to the Discussion:

> 'It is theoretically possible that a sufficiently complex RL model could be tuned to show a plateau-like component of RPE at intermediate levels of training. Kim et al.[50] showed how temporal discounting can produce upwards ramping RPE responses that resemble ramping dopamine responses that have been reported[10,32,43,44,59-61]. However, such upward ramps were rarely observed in our data. In RL models that endow the agent with

> a fine sense of the passage of time, such that each time point can be represented as a distinct state, it is also possible to find a small hump in the RPE signal in between the cue and reward delivery that becomes progressively earlier in every trial[1,40]. This hump only occurred in the middle of training, not at the start or end. If the right range of mid-training trials were analyzed, a combination of an upward ramp due to temporal discounting and a moving hump due to progressive transfer of the RPE signal to earlier states in trials could potentially add up to a roughly plateau-like response. We observed plateaus at the end of training, but because we did not keep training our mice after they reached criterion, it is impossible to say based on the present experiments whether the plateaus we observed would persist indefinitely with additional training. Such humps occur at the expense of the reward response, and we did not observe a diminution of the reward response when the plateau components arose. Another potential mechanism that can add a hump to the RPE signal is uncertainty as to exactly when the reward will be delivered[51,62]. In our task, the animal received a very clear reward delivery signal (i.e., the extinguishment of the rewarded cue), so this effect is unlikely to play a role in the present study. Also, a hump due to uncertainty will necessarily be close to the actual reward delivery time, unless the animal systematically overestimates the passage of time between cue and reward. Extensive additional modeling work will be required to determine whether sufficiently complex RL models actually can produce plateau-like RPE signals between cue and reward, and such models might have so many free parameters that they could be fitted to arbitrary data. We thus did not pursue these questions here.’

and probabilistic tasks (i.e. decrease of reward responses and increase of cue responses over training and inhibition by reward omission, monotonic activation modulation by reward probability),

We again thank the Reviewer. We have made the following changes to the text to highlight the specific aspects of responses to that do support the RPE interpretation and indication of future studies.

> 'There was a substantial increase in the cue response with learning, consistent with an RPE signal, but only a minor decrease in outcome response, in contrast to the large decrease expected on the basis of RPE algorithms.'
>
> '**Single-cue probabilistic reward exhibited some RPE-like features**' (added as a paragraph header)
>
> 'These response patterns, along with the more medial region’s reward response early but not late in cue reversal training, were unique in the present data set in being consistent with an RPE interpretation.'

with some differences between DMS and DLS, which is consistent with a previous study (Parker et al., 2016, Tsutsui-Kimura et a., 2020) and contradicts with authors’ main claim.

We apologize for being so unclear about what our main claim is, with the hope that the choice of wording in the title would imply only that there are certain features of dopamine signals that contrast with RPE signals, not that there are no features at all that are consistent with the RPE interpretation. To help clarify this point, we have modified the Abstract to read:

> 'Dopamine release responses differed for the medial and lateral sites. In neither sector could these be accounted for by classic reinforcement learning alone as classically applied to the activity of nigral dopamine-containing neurons.'

3. Related to 2, while authors contrast their observation with TD errors, there is no formal prediction or no formal test.

We have implemented a simple Q-learning model for comparison of RPE signals to our dopamine recordings, and made the comparison in the new Supplementary Figure 3 and accompanying text:

> **‘Supplementary Fig. 3**. **Comparison of model RPE results with experimental results for dopamine release responses.** See Supplementary Text. **a** State transition diagram for the Q-learning model. "S1" - "S5": model states. Each arrow exiting from each state is annotated with the action represented and its net reward value. **b-d** First column shows trial-averaged RPE values from the Q-learning model; second and third columns are reproduced from main **Fig. 2a** for convenience of comparison. **b** Random reward responses. The model was trained only on the random reward task. Model trials were aligned on reward delivery (time 0) and averaged for comparison to the trial-averaged dopamine responses. First licks normally occurred on the next time step. Blue: first 3 trials; red: last 10 trials. **c** Cue discrimination training. Training began directly with the cue discrimination task. Model trials were aligned on cue onset (time 0). Cue termination and reward delivery were 1.5 sec after cue onset. Blue: first 10 trials; red: last 10 trials. Note that the discrimination task plots copied from **Fig. 2a** show cue onset at time 0.5. **d** Reversal discrimination training, aligned as in b. Reversal training followed cue discrimination training in order to show the reversal dynamics. Blue: first 10 trials; red: last 10 trials.’
>
> '**Simple Q-learning RL model cannot account for our observations**
> It was clear that the absence laterally of transfer of transient dopamine from outcome to predictive cue and absence medially of a positive outcome response were not in accord with classic RPE models[49]. But this was not so clear for the plateau responses. To assess this issue, we constructed a simple Q-learning model. As expected, this simple model soundly replicated the well-known RPE signals. However, it did not replicate plateau responses, especially the fact that they emerged when the second, unrewarded cue light has added (**Supplementary Fig. 3**). Kim et al.[50] have found that the prolonged ramping dopamine signals reported experimentally[32] can represent RPE signals when temporal discounting ("discount rate" in Ref 2) is a factor. In our data, plateau responses varied in shape from trial to trial (**Supplementary Fig. 3**), but in the aggregate and in most individual mice they were more flattened than ramp-like, and in some animals, even resembled reversed (decreasing) ramps, indicating that temporal discounting was not responsible. Our RL model also did not account for why plateau responses were absent in the cue association task and abruptly emerged when the alternative but never simultaneously presented cue was presented in the sessions (**Supplementary Fig. 3**). As a test for whether the order of task paradigms could be important for these responses, we switched tasks in the Q-learning model so that the cue discrimination preceded rather than followed the cue association sessions. The model did not behave as the mice did: it gave transients at both cue and outcome. More stringently, any RPE model is strongly challenged to explain the decrease of cue-associated phasic response as mice learn the task, or the increase of reward-evoked responses in random reward paradigm, even if they incorporate high eligibility trace (lambda) and sensory uncertainty[50,51]. Future studies should address how these issues could be addressed in more complex models that might account for the behavior and dopamine signals reported here.'

For example, if they want to claim for (or against) transfer of signals from outcome periods to cue periods, they should quantify the time-course of those signals at the trial basis in single animals. Because of complication of history until cue discrimination (see above), the reversal task would be a proper task for such quantification in this dataset. Basically, any claim should be based on proper quantification, which is largely lacking in the manuscript.

We truly wish we could quantitatively analyze individual trials as the reviewer wisely suggests. Unfortunately, dopamine concentration in the dorsal striatum changes with time in much too bursty a fashion to make single trial analysis practicable, in both the present photometry recordings and our previous FSCV recordings (Schwerdt et al., 2017). We show examples of this in new Figure 3, where we also defend as best we can our use of averaged data in place of single trials. Our comparison of RPE signals from our new Q-learning model in new Supplementary Fig. 3 shows such dramatic qualitative differences that it would be difficult to measure them quantitatively. These are described in the new Supplementary Text.

4. “Plateau response” at cue is sustained only for 1.5s, since there is no delay from cue and outcome in this task. Considering these slow dopamine sensors, this activity is likely the second peak that previous studies observed in dopamine neurons: first detection, and then value (Nomoto et al., 2010). Authors should have a long delay period from a cue to an outcome or examine electric spikes to determine whether the activity plateaus.

We appreciate the Reviewer's comment. It is true that the plateaus may not be originated from the sustained electrical activities of dopamine axons. However, in either case, our claims stand. Given that the plateaus are not solely the decay of cue responses (they are absent in the cue association task), there are activities between the cue and reward presentations that cannot be explained as the transfer of activations of RPE model. We thus discussed this point as a caveat of using slow dopamine sensors.

> 'We are aware of caveats that should accompany our conclusions. The tasks were variants of Pavlovian tasks and lacked the richness of much behavioral learning, decision-making and response variety. We used a fixed sequence of paradigms across animals as representative of the many switches that can occur in daily experience, but we are aware that the results could be constrained by this training sequence. We used both D1R-based (i.e., dLight1.3b and GRAB$_{DA3m}$) and D2R-based (i.e., GRAB$_{DA2m}$) dopamine sensors. Decay time constants for GRAB$_{DA2m}$ and GRAB$_{DA3m}$ are, respectively, 1.3 sec[48] and ~600 msec[66], but that for dLight1.3b has not been determined. This imposes a lower temporal resolution on our data as compared to electrical recordings. We only sampled relatively restricted parts of more medial and lateral parts of the centrodorsal striatum, and did not consider the compartmentalization of the striatum, in which striosome and matrix compartments have different relationships to dopamine[33,35,67]. We used photometry, a recording method that measures the local sum of extracellular dopamine, whereas dopamine likely works both at individual synapses[68] and as an ambient non-synaptic modulator. Our findings cannot address the synaptic actions of dopamine because our measurements are probably dominated by extrasynaptic dopamine. RPE-observing dopamine signaling might instruct reinforcement plasticity only in a small subset of synapses that convey the relevant information, as suggested by reports of multiple, multiplexed responses of dopamine and dopamine neuron firing[10,43,44,59-61]. Despite these uncertainties, the surprises that emerged in our experiments open new opportunities to probe and to model mechanisms underlying striatum-based learning and its modulation by dopamine.'

In addition, PCA is indirect to quantify the sustained activity. The activity level during later periods should be directly used.

The middle panel of the bottom row in Figure 2a illustrates direct quantification using the area under the curve value during the final 500-msec period. Additionally, we include a PCA map to visualize the anatomical distribution of plateau patterns. Also, Figure 5a and b show the extraction by means of PCA of exactly the aspect of the plateaus that we were most interested in, i.e., the abrupt rise and extremely slow decay. By definition, PC1 identifies the most common source of variance across the waveforms. We tentatively interpret the small humps and other departures from the shape of PC1 as being due to some other, less predictable source. We revised the text to clarify this point. We have made this addition to the Results section:

> 'Because PCs account for correlated components of variance across all dimensions (i.e., all time points, in this case) it is tempting to assume that each PC ultimately identifies a single causal source of variance. This is not a logically necessary inference, but it is generally difficult to find an alternative explanation for how two or more sources of variance can be correlated unless they do in fact share some ultimate common cause. We thus tentatively interpret PC1 as reflecting an input to dopamine release that is relevant to the majority of mice, and the other PCs as representing factors that may be relevant in smaller numbers of cases.'

5. Unique responses to visual stimuli in dopamine in DMS had been reported (Moss et al., 2021). Thus, the initial quick cue responses are likely to be sensory (or detection), and may not be subject to value transfer, but instead may decay with familiarity.

We thank the Reviewer for mentioning this paper on this very important issue. Sensory responses are one of many factors in addition to RPE that we believe are likely to be important to dopamine levels. Moreover, in our data, both cue response and plateaus scale with values as shown in Fig. 2a (cue discrimination and reversal tasks) and Fig. 3a and b (probabilistic reward task). In other words, both cue response and plateaus are not entirely determined by sensory stimuli but also depend on the values. We discussed this point in the 'Dynamic shaping of striatal dopamine release responses during learning' section. In previous studies, it has been shown that the initial dopamine response encodes sensory stimuli (Nomoto et al., 2010, Moss et al., 2021), rather than the values in RL. To formulate the model to explain the dopamine responses shown here, sensory as well as value responses will have to be incorporated in future studies.

Authors should have controls without any outcomes (single cue, no outcome).

We agree with the Reviewer's important comment. Trial-and-animal-averaged results for unrewarded trials can be seen in Fig. 2a; trial-averaged results for unrewarded trials in one animal are shown in Fig. 2c and d; and trial-and-animal-averaged results for unrewarded trials grouped by behavioral performance level can be found in Fig. 3e.

6. Previous studies reported contra-lateral bias in dopamine signals, especially in DMS (Moss et al, 2021). Responses to visual cues contra- vs ipsi-lateral to the recording site, and to associated outcomes should be analyzed separately.

We found no statistically significant differences between cue responses based on laterality, and have added Supplementary Fig. 2 and the following text in the Results section:

> 'We also found no differences between ipsilaterally and contralaterally presented cues (**Supplementary Fig. 2**) and merged these as well.'

7. Inhibition of dopamine activity by licks in initial stages of training had been reported with specific interpretation (Coddington and Dudman, 2018). Authors should test the original authors' model to be a follow-up study.

This is an excellent suggestion; we thank the Reviewer and will corporate in the future computational study that we propose to do. A full computational model will definitely include this.

8. Difference of activity patterns between DMS and DLS should be quantified for each claim.Because sex and performance dramatically affected dopamine activity in this study, authors should consider those together instead of testing one-by-one, to determine contribution of location.

The Reviewer makes a good point, and we agree that this would make the paper more complete. Our primary focus was in looking for qualitative changes in the overall patterns of dopamine release, which are hard to quantify, and it happened that the quantitative AUC analyses we show were easy to do as part of that process. Unfortunately, it is difficult for purely organizational reasons for us to expand the scope of the AUC analyses at this time.

9. Half of sessions in the probabilistic reward task were excluded and "the session with the better recording quality was selected for analysis". Why do not they just average both sessions? How was the "recording quality" in other tasks?

We thank the reviewer to point this out. The dopamine sensors suffer from photobleaching after prolonged chronic recording, and the fiber tips may get fouled as well, so in the probabilistic and extinction sessions at the end of the recording series for a given animal, the signal sometimes became weak. We therefore analyzed only the probabilistic session with the greater variation in fluorescence level. We also compared the range of fluorescence levels in the last session of reversal discrimination to those in the previous session to make sure there was no loss of signal. This was not a problem in the earlier sessions.

We now have added this to the revised manuscript by saying:

> 'As the weeks of training accumulated, some sensors in some mice suffered from bleaching and/or fouling of the sensor tips. Our yields were accordingly reduced.'

More comments (not minor)(!)

We do not take any of the Reviewer's comments as minor. Each point is valuable and is addressed carefully.

10. Why are signals even before a cue always synchronized on different days or in different locations such as in Figure 2a column2 and Figure 3a, c, d, e?

If the Reviewer is referring to the low amplitude reversed ringing that precedes the cue presentation in those figures, that would be due to the application of a zero-phase lowpass filter during the analysis. We have inserted a note to that effect in the legend to Figure 2a:

> 'The small oscillations preceding cue onset in some plots are bandpass filtering artifacts.'

In the case of Fig. 2a, as is stated in the figure legend,

> 'Data from random reward sessions, which included those inserted late in training, are aligned with reward delivery at t = 1.0 sec.'

and so the activity before the event marker most likely reflects preparatory activity to start licking in response to the delivery of the reward, which may have been detectable by the animal in spite of there being no cue.

11. Related to 2, dynamical changes of lick patterns (during ITI and cue periods) in each trial type in each task across days should be shown to verify proper learning.

We agree with the Reviewer. Ideally, we would like to show many parameters on a day-by-day basis for each animal, but this would lead to a very long and complex paper beyond the scope of this initial report. We chose instead to use the auROC-based learning criterion specified in the Methods subsection titled "Cue discrimination and reversal discrimination training" to avoid presenting too much detail.

12. Learning progress should be summarized to show how many animals progressed to the next level and how long animals spent for each stage of learning.

We have now added a new Supplementary Table 1, which shows this information.

13. Major findings should be verified using GFP expression (or GRAB-DA-mut, see Costa et al., 2023) as controls for motion artifacts in each area with the same normalization with 405w light.

We used an isosbestic control recording channel for this purpose, as is customary in published photometry papers.

14. Average activity patterns in all the figures should have error bars, instead of having only averages.

Figures 2, 3, and 6 have all now been revised accordingly.

15. Z-score should be calculated using only ITI to compare activity across sessions. Show all trials in an example animal to verify success of normalization.

There are several different types of baseline measures that are commonly used in the published literature, of which ITI activity is one. The one that we generally prefer, which we chose to use here, is the whole data recording for the entire session, including both ITI and task activity. We have additionally clarified the text by inserting this new sentence in the Methods section:

> 'Z-scores were computed using the mean and standard deviation calculated for the whole data recording for the entire session, including both inter-trial interval and task activity.'

We have also added the new Figure 3, which shows all trials for a single mouse as an example.

16. Task structures such as ITI distribution (uniform vs exponential) and trial types (% free water) should be clearly written.

We have modified the text in the Methods that read "trials with randomly varying trial durations of 8-48 sec" to read:

> 'trials with uniformly distributed random durations of 8-48 sec.'

17. Are Figure 2b-d from lateral or medial DS? Both should be shown separately.

We have modified the legend for Figure 2b-d to read:

> '**b, c** Transition from initial cue-association (**b**) to cue discrimination (**c**) training shown for a single mouse (pa96, medial). In **c**, dopamine response in rewarded (top) and non-rewarded trials are shown. Vertical lines indicate cue onset. Dopamine traces aligned as in **Fig. 1d** and **e** illustrate responses in consecutive sessions from last cue conditioning (**b**, bottom) to first three cue discrimination sessions (**c**, right), illustrating gradual emergence of prolonged plateau dopamine release during cue presentation period in rewarded trials of cue discrimination sessions. Color scale shows z-scores of dF/F, which ranged from −1 to +4. Color-coded line plots on right. **d** Superimposed session-averaged dopamine release in response to rewarded and non-rewarded cue onsets recorded in a mouse (animal pb43, medial) during all training sessions, from day 1 (light blue) to criterion (dark blue), of cue (2 left panels) and reversal (2 right panels) discrimination training. Shaded purple and pink boxes indicate 1.5-sec cue period.'

Additional References

Moss, M.M., Zatka-Haas, P., Harris, K.D., Carandini, M., and Lak, A. (2021) Dopamine axons in dorsal striatum encode contralateral visual stimuli and choices. *J. Neurosci.*, 41: 7197-7205.

Nomoto, K., Schultz, W., Watanabe, T., and Sakagami, M. (2010) Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. *J. Neurosci.*, 30: 10692-10702.

Schwerdt, H.N., Shimazu, H., Amemori, K.-I., Amemori, S., Tierney, P.L., Gibson, D.J., Hong, S., Yoshida, T., Langer, R., Cima, M.J., and Graybiel, A.M. (2017) Long-term dopamine neurochemical monitoring in primates. *Proc. Natl. Acad. Sci. U S A*, 114:13260-13265.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This is an important work on one of the most studied ideas in brain research, and one that has penetrated popular culture, the idea that dopamine release signals reward prediction error. The paper argues against this notion, and offers evidence that it cannot be correct in its strongest form.

The authors have revised the manuscript, and the revised version address all of the issues I had with the original. This is an exciting paper that will have substantial impact.

Reviewer #2 (Remarks to the Author):

The authors have done an excellent job addressing my comments. While I am curious about some possibilities that are not fully worked out here, I am onboard with the authors' approach of discussing outstanding questions and the need for further investigations for additional clarity. Overall, I think the authors have done enough for a solid publication and I am okay with the manuscript being published as is.

Reviewer #2 (Remarks on code availability):

I did not read the code line by line, but the structure seems good and largely self-explanatory.

Reviewer #3 (Remarks to the Author):

The authors conclude that, “at least at the population level that can be imaged by fiber photometry, dorsal striatal dopamine release responses do not fully follow RPE formulations in either more medial or more lateral regions, but exhibit instead unpredicted heterogeneities across striatal districts.” The results are noteworthy and significant for the advancement of the basal ganglia research.

The revised manuscript addressed most of minor issues this reviewer pointed out. However one issue is still unclear to me.

“dopamine release responses exhibit heterogeneities across striatal districts” is one of the main conclusions of this paper. However the descriptions of districts are frustratingly difficult to understand. Use of various descriptions for striatal locations distract reading. “dorsal striatum”, “caudoputamen”, medial or more lateral regions”, “centromedial” and “centrolateral” appear first time in Discussion, “central-medial sites and central-lateral” also appear first time in Discussion. Most of the descriptions of the Results are medial vs lateral regions of the dorsal striatum. No clear descriptions of heterogeneities of antero-caudal nor dorso-ventral striatum are given.

“standardized coordinate" is still difficult to understand for me. How the A-P position defined in “standardized coordinate" system? “A-P coordinates were determined by comparing histological sections to the atlas and are given relative to bregma as in the atlas.” Is this means A-P positions are in mm? Why A-P scale in Fig. 6e and 6g differ so much? How M-L positions in 6g and 6h can exceed 1? Why M-L scale in Fig. 6c an 6g differ so much? Which site is anterior? Which site is medial? Figure 6 introduces data in A-P and D-V dimensions or various “striatal districts” which are very difficult to understand.

Line 838: “Fig. 1b and f, M-L coordinates are our standardized coordinate ratio” The values appear not derived from “the standardized medial-lateral coordinate of the probe tip was calculated as (x5 – x1) / (x2 – x1),”

Reviewer #4 (Remarks to the Author):

Authors addressed many previous concerns by formally constructing a counter model. There are still major technical concerns.

1. About motion artifacts.

13. Major findings should be verified using GFP expression (or GRAB-DA-mut, see Costa et al., 2023) as controls for motion artifacts in each area with the same normalization with 405w light.

We used an isosbestic control recording channel for this purpose, as is customary in published photometry papers.

Isosbestic signals have been used as a control, especially for photometry recording with GCaMP. However, it does not guarantee that the system can eliminate motion artifacts unless it is tested in each brain area in each study. Specifically, Isosbestic points and the effectiveness of those signals have not been established for recording with dopamine sensors (see Labouesse et al., 2020). For example, the isosbestic point for DA2h is 440 nm (Sun et al., 2020), and dLight2 is insensitive to 405 nm excitation light in vivo (Robinson et al., 2019). Thus, it is important to perform control experiments to confirm at least the core findings. See also Siciliano and Tye, 2019 for importance of control experiments to evaluate motion artifacts.

2. About normalization.

15. Z-score should be calculated using only ITI to compare activity across sessions.

There are several different types of baseline measures that are commonly used in the published literature, of which ITI activity is one. The one that we generally prefer, which we chose to use here, is the whole data recording for the entire session, including both ITI and task activity. We have additionally clarified the text by inserting this new sentence in the Methods section:

If the signals are normalized with the time window including those signals, it is hard to compare results across sessions. For example, anti-correlation between cue responses and reward responses in some cases might be explained solely by normalization. This is why it is important to confirm the main findings with different way of normalization.

Replies to Reviewer’s Comments

Reviewer #1 (Remarks to the Author):

This is an important work on one of the most studied ideas in brain research, and one that has penetrated popular culture, the idea that dopamine release signals reward prediction error. The paper argues against this notion, and offers evidence that it cannot be correct in its strongest form.

The authors have revised the manuscript, and the revised version address all of the issues I had with the original. This is an exciting paper that will have substantial impact.

We thank the Reviewer for these very positive and encouraging remarks.

Reviewer #2 (Remarks to the Author):

The authors have done an excellent job addressing my comments. While I am curious about some possibilities that are not fully worked out here, I am onboard with the authors' approach of discussing outstanding questions and the need for further investigations for additional clarity. Overall, I think the authors have done enough for a solid publication and I am okay with the manuscript being published as is.

Reviewer #2 (Remarks on code availability):

I did not read the code line by line, but the structure seems good and largely self-explanatory.

We thank the Reviewer for this favorable code review.

Reviewer #3 (Remarks to the Author):

We thank this Reviewer for very helpful editing of our manuscript and for excellent suggestions and critiques. We have incorporated all of the Reviewer’s suggestions within our revised manuscript, which is better for these changes.

The authors conclude that, “at least at the population level that can be imaged by fiber photometry, dorsal striatal dopamine release responses do not fully follow RPE formulations in either more medial or more lateral regions, but exhibit instead unpredicted heterogeneities across striatal districts.” The results are noteworthy and significant for the advancement of the basal ganglia research.

We thank the Reviewer for these positive summary remarks.

The revised manuscript addressed most of minor issues this reviewer pointed out. However one issue is still unclear to me.

“dopamine release responses exhibit heterogeneities across striatal districts” is one of the main conclusions of this paper. However the descriptions of districts are frustratingly difficult to understand. Use of various descriptions for striatal locations distract reading. “dorsal striatum”, “caudoputamen”, medial or more lateral regions”, “centromedial” and “centrolateral” appear first time in Discussion, “central-medial sites and central-lateral” also appear first time in Discussion. Most of the descriptions of the Results are medial vs lateral regions of the dorsal striatum. No clear descriptions of heterogeneities of antero-caudal nor dorso-ventral striatum are given.

We have carefully examined our manuscript, and we found several ambiguities and/or errors in regard to this point. We apologize and have made the following corrections.

'Histological localization of recording sites' section: minor re-organization, and corrected text to read:

> 'In Fig. 1b and f, M-L coordinates are calculated in this way. The D-V coordinates shown in these figures are given by the distance (mm) from dorsal surface of the striatum as it appears in the same section containing the probe track.'

The X axis and Y axis labels in Fig. 6g and h were reversed. This has now been corrected.

Supp. Fig. 1 had the same axis label switch, and the numbers on the X axis were incorrect. This has now been corrected.

We have also adopted a uniform terminology for referring to striatal regions, including "dorsal striatum" in place almost entirely of the rodent-specific "caudoputamen", and the phrases "centromedial" and "centrolateral" to designate the locations of the medial half and lateral half of our set of recording sites respectively. This definition is now explicitly articulated in the Introduction:

> 'We found shifts in the patterning of dopamine release signals as successive versions of the cue-association tasks were acquired, and sharp differences in the dopamine release patterns between the centromedial and centrolateral striatal sites from which we recorded in the 67 mice sampled.'

Finally, we substantially rewrote the majority of the "Histological localization of recording sites" subsection of the Methods, including the following new or revised text:

> 'The dorsal striatum contains many semi-independent histochemical and physiological gradients in all three cardinal dimensions. The boundaries of striosomes can be sharply delineated with some histochemical stains, but there are no known similarly clear boundaries subdividing the dorsal striatum at maturity into districts at a larger scale. In placing our probes, we avoided the medial and lateral extremes of the caudoputamen and thus populated roughly the central half of the volume of the striatum with probe tips.

> We defined a standardized coordinate system within the striatum to accommodate its slightly irregular shape, as follows.'
>
> ' We refer to this coordinate system as "relative position" or "standardized coordinate" in the coronal plane (shown in **Fig. 6 c** and **d** and **Supplementary Fig. 8**). A-P coordinates (specified in mm) were determined by comparing histological sections to the atlas[69] and are given relative to bregma as in the atlas.'
>
> 'A slightly different coordinate system was used by a second person to classify probes as "centromedial" or "centrolateral". Recording sites were classified as "centromedial" if the distance from the midline to the tip of the probe was less than 0.6 (60%) of the mediolateral distance from the midline to the lateral edge of the striatum in the coronal section containing the site. In **Fig. 1b** and **f**, M-L coordinates are calculated in this way. The D-V coordinates shown in these figures are given by the distance (mm) from dorsal surface of the striatum as it appears in the same section containing the probe track. These measurements differ from the "standardized coordinate" system described above. Depending on the A-P plane of section, the division between "centromedial" and "centrolateral" corresponded to around 0.4 to 0.5 M-L in the "standardized coordinate" system.'

There follow a number of specific questions from the Reviewer, which we answer one at a time here:

“standardized coordinate" is still difficult to understand for me. How the A-P position defined in “standardized coordinate" system?

We have added this text to the Methods:

> 'A-P coordinates (specified in mm) were determined by comparing histological sections to the atlas[69] and are given relative to bregma as in the atlas.'

“A-P coordinates were determined by comparing histological sections to the atlas and are given relative to bregma as in the atlas.” Is this means A-P positions are in mm?

Yes.

Why A-P scale in Fig. 6e and 6g differ so much? How M-L positions in 6g and 6h can exceed 1?

This was due to an axis labeling error, and it has now been corrected.

Why M-L scale in Fig. 6c an 6g differ so much?

This was also due to an axis labeling error. We have now corrected them, thanks to this Reviewer.

Which site is anterior?

We have added labels to Fig. 6c, e, g to indicate which end is which.

Which site is medial?

We have added labels to Fig. 6c, e, g to indicate which end is which.

Figure 6 introduces data in A-P and D-V dimensions or various “striatal districts” which are very difficult to understand.

The concluding sentences of the first paragraph of the Discussion now read:

> 'We conclude that, at least at the population level that can be imaged by fiber photometry, dorsal striatal dopamine release responses do not fully follow RPE formulations in either more medial or more lateral regions, but exhibit instead unpredicted heterogeneities that we have shown in detail along the mediolateral dimension, and indicated briefly in **Fig. 6** in the other two dimensions. These findings encourage further work on how the multitudes of striatal circuits are coordinated to instruct learning and to modulate behavior under the influence of dopamine.'

We hope this is sufficiently clear.

Line 838: “Fig. 1b and f, M-L coordinates are our standardized coordinate ratio” The values appear not derived from “the standardized medial-lateral coordinate of the probe tip was calculated as (x5 – x1) / (x2 – x1),”

We again thank the Reviewer for finding this extremely confusing error. As noted above, this sentence has been moved and corrected to read as follows, where 'in this way' refers to the previous sentence, 'Recording sites were classified as "medial" if the distance from the midline to the tip of the probe was less than 0.6 of the distance from the midline to the lateral edge of the striatum':

> 'In **Fig. 1b** and **f**, M-L coordinates are calculated in this way. The D-V coordinates shown in these figures are given by the distance (mm) from dorsal surface of the striatum as it appears in the same section containing the probe track.'

Reviewer #4 (Remarks to the Author):

Authors addressed many previous concerns by formally constructing a counter model. There are still major technical concerns.

We are glad that the Reviewer was satisfied with our replies to his/her questions that we answered, and we thank him/her for clarifying the remaining issues. We have gone to extraordinary lengths to address the Reviewer’s new criticisms. For example, we have experimented with new mice, ordered new reagents, performed multiple experiments, added accelerometers to our setups, and developed modeling to join the data presentation. We most respectfully request that we now be allowed to publish these valuable data and thank the Reviewer.

1. About motion artifacts.

13. Major findings should be verified using GFP expression (or GRAB-DA-mut, see Costa et al., 2023) as controls for motion artifacts in each area with the same normalization with 405w light.

We have performed experiments to address this issue.

Unfortunately, in spite of sincere efforts, we have not been able to find the 2023 paper with Rui M. Costa as first author (we assume that is the correct "Costa"). There is a topically relevant paper from Costa lab that was accepted in 2023 and published in the "15 February 2024" issue:

> Tang, J.C.Y., Paixao, V., Carvalho, F. et al. Dynamic behaviour restructuring mediates dopamine-dependent credit assignment. Nature 626, 583–592 (2024). https://doi.org/10.1038/s41586-023-06941-5

However, the Tang et al. paper does not describe use of GFP or isosbestic excitation as controls for motion artifacts. Nonetheless, we have performed an extensive additional series of control experiments and analyses to more completely address the issue of motion artifacts, which are described below.

Isosbestic signals have been used as a control, especially for photometry recording with GCaMP. However, we point out that it does not guarantee that the system can eliminate motion artifacts unless it is tested in each brain area in each study. Specifically, Isosbestic points and the effectiveness of those signals have not been established for recording with dopamine sensors (see Labouesse et al., 2020). For example, the isosbestic point for DA2h is 440 nm (Sun et al., 2020), and dLight2 is insensitive to 405 nm excitation light in vivo (Robinson et al., 2019). Thus, it is important to perform control experiments to confirm at least the core findings. See also Siciliano and Tye, 2019 for importance of control experiments to evaluate motion artifacts.

We thank the Reviewer for clarifying these issues regarding controls for motion artifacts.

We have now run control experiments using GRABDA-mut and an accelerometer to control very specifically for motion artifacts, and we have also done additional analyses of our existing data to confirm that our putative DA signals cannot be ascribed to the limitations of our isosbestic control methods. These are described in detail in a new Supplementary Information section, "Controlling for Motion Artifacts". Our data are confirmed by these extensive and time-costing analyses.

2. About normalization.

15. Z-score should be calculated using only ITI to compare activity across sessions.

We thank the Reviewer for this comment. We have now done this for our Fig. 2d, in which we compare activity across sessions. This is shown in the new Supplementary Information section, "Signal Processing Pipeline and Alternative Z-Scoring Method", subsection "Reference Data Selection for Z-Scoring", along with results that do not use Z-scoring at all. The results were not easily distinguishable from our original results.

If the signals are normalized with the time window including those signals, it is hard to compare results across sessions. For example, anti-correlation between cue responses and reward responses in some cases might be explained solely by normalization. This is why it is important to confirm the main findings with different way of normalization.

We thank the Reviewer for this observation, and we took it seriously. The new Supplementary Information section, "Signal Processing Pipeline and Alternative Z-Scoring Method", compares our results with analogous results obtained using strictly ITI-only-based z-scoring, and shows the transformation from the raw recorded fluorescence signals through each step of our analysis pipeline. We were not able to find any indications that our choice of analysis methods qualitatively altered the results.

REVIEWERS' COMMENTS

Reviewer #3 (Remarks to the Author):

The revised manuscript addressed all the issues this reviewer pointed out. The results are noteworthy and significant for the advancement of the basal ganglia research. I think the manuscript is ready for a publication.

Reviewer #4 (Remarks to the Author):

Authors addressed all previous concerns.

REVIEWER COMMENTS

Reviewer #3 (Remarks to the Author):

The revised manuscript addressed all the issues this reviewer pointed out. The results are noteworthy and significant for the advancement of the basal ganglia research. I think the manuscript is ready for a publication.

Reviewer #4 (Remarks to the Author):

Authors addressed all previous concerns.

We thank the Reviewers for their helpful comments and suggestions throughout the review process. Their advice made the result reporting much clearer and the manuscript much stronger.