nature neuroscience

Article

https://doi.org/10.1038/s41593-023-01310-x

Dopaminergic prediction errors in the ventral tegmental area reflect a multithreaded predictive model

Received: 12 November 2021

Accepted: 16 March 2023

Published online: 20 April 2023

Check for updates

Yuji K. Takahashi[1], Thomas A. Stalnaker[1], Lauren E. Mueller[1], Sevan K. Harootonian[2], Angela J. Langdon[3,4] & Geoffrey Schoenbaum[1,4]

Dopamine neuron activity is tied to the prediction error in temporal difference reinforcement learning models. These models make significant simplifying assumptions, particularly with regard to the structure of the predictions fed into the dopamine neurons, which consist of a single chain of timepoint states. Although this predictive structure can explain error signals observed in many studies, it cannot cope with settings where subjects might infer multiple independent events and outcomes. In the present study, we recorded dopamine neurons in the ventral tegmental area in such a setting to test the validity of the single-stream assumption. Rats were trained in an odor-based choice task, in which the timing and identity of one of several rewards delivered in each trial changed across trial blocks. This design revealed an error signaling pattern that requires the dopamine neurons to access and update multiple independent predictive streams reflecting the subject's belief about timing and potentially unique identities of expected rewards.

Evidence has tied phasic activity in dopamine neurons to the prediction error in temporal difference reinforcement learning (TDRL) models[1–5]. Yet these models make significant simplifying assumptions, particularly with regard to the structure of the predictions fed into the hypothesized TDRL comparators—the dopamine neurons—which usually consist of estimates of the current and future scalar values of a single chain of timepoint states. Although this predictive structure has been sufficient to explain the error signals observed in many studies[6–14], it can fall short under a variety of conditions. One early example of such a condition occurred in studies in which the timing of the reward varied. The original TDRL implementations predicted that a delayed reward appearing earlier than expected should result in suppressed activity—a negative prediction error—later in the trial when the reward would normally have been expected. This was not observed in the data[7,15,16]; instead, when a delayed reward appeared early, it induced an increase in firing when it occurred but no suppression at its later 'omission'. Thus, the biological system reacted as if the earlier appearance of the reward somehow predicted its later omission, that is, that the animal realized the reward had arrived early and did not expect a second reward. To account for this, a model was introduced that incorporated a reset in the prediction mechanism triggered by the terminal event ending the trial—the appearance of the reward[17]. More recent models attributed the apparent reset of reward expectations to changes in belief about the hidden state of a task, allowing a transition to the intertrial interval to be inferred after delivery of a variably timed reward[18–20]; although intuitive, this 'fix' does not address what governs the identification of the terminal event, especially in more naturalistic settings in which there would be no clear termination of a series of potential rewards, or in which some rewards in the series might differ in features besides their timing. In the present study, we investigated this question, taking as our starting point the observation that changes in reward timing seem to reset

[1]Intramural Research Program, National Institute on Drug Abuse, Baltimore, MD, USA. [2]Psychology Department, Princeton University, Princeton, NJ, USA. [3]Intramural Research Program, National Institute of Mental Health, Bethesda, MD, USA. [4]These authors jointly supervised this work: Angela J. Langdon, Geoffrey Schoenbaum. e-mail: yuji.takahashi@nih.gov; angela.langdon@nih.gov; geoffrey.schoenbaum@nih.gov

https://doi.org/10.1038/s41593-023-01310-x

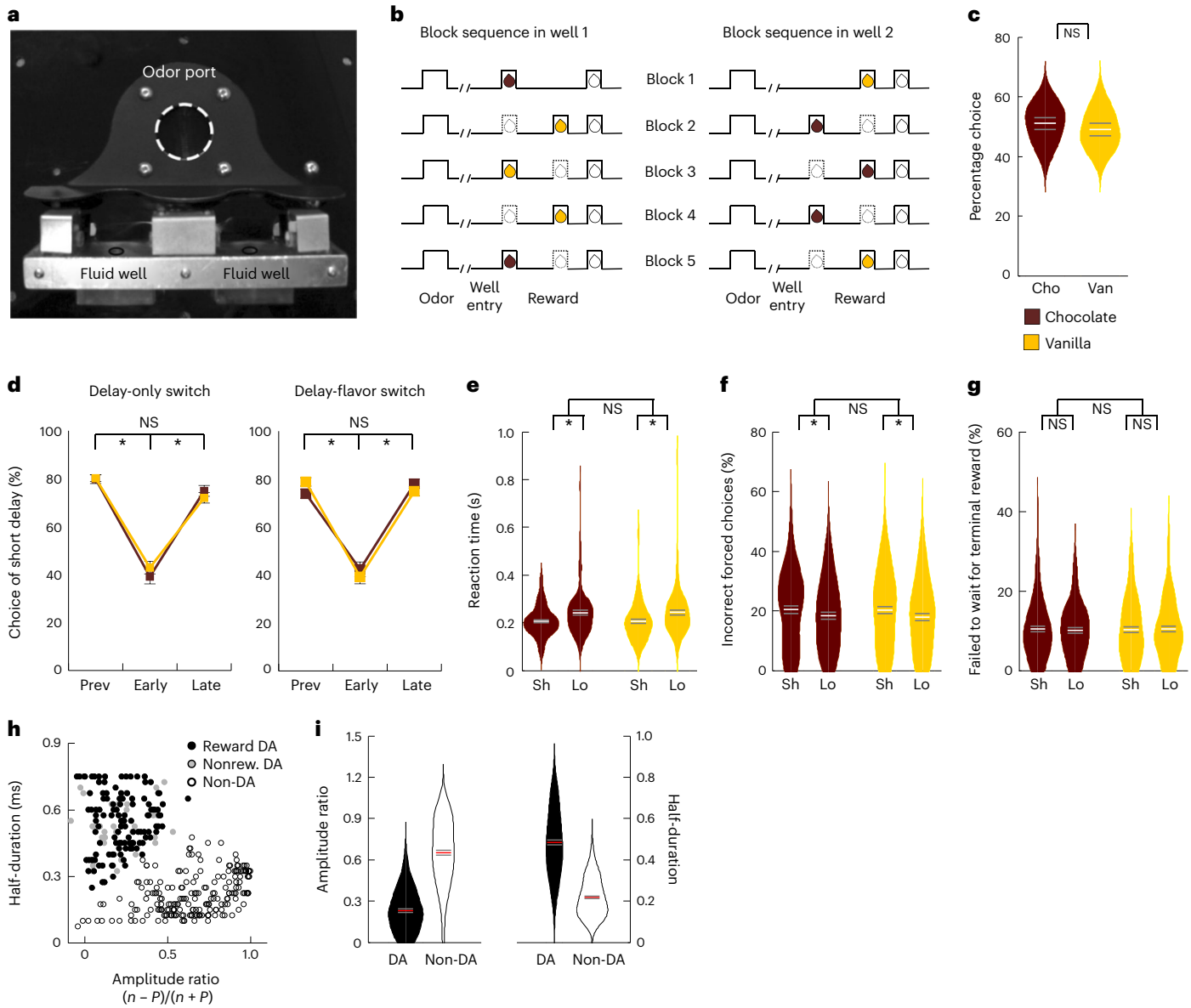


Fig. 1 | Task design, behavior and identification of putative dopamine neurons. a, Picture of apparatus used in the task, showing the odor port (~2.5 cm in diameter) and two fluid wells. **b**, Deflections indicating the time course of stimuli (odor and reward) presented to the animal on each trial. The dashed lines show when a reward was omitted and the solid lines when a reward was delivered. At the start of each recording session, one well was randomly designated to deliver a single drop of flavored milk (chocolate or vanilla) after a short delay (0.5 s). In the other well, one drop of the other-flavored milk was delivered after a long delay (1.0- and 2.0-s delay on first 2 trials, respectively, and 3.0-s delay thereafter). In both wells, a single drop of water—the terminal reward—was also delivered at a 5.0-s delay from well entry. In the second and fifth blocks, both delay and flavors were switched (delay-flavor switch) compared with the previous block. In the third and fourth blocks, the delay was switched without changing the flavors (delay-only switch). **c**, Chocolat (Cho)- and vanilla (Van)-flavored milk shown as equally preferred in two-flavor choice tests (one-way analysis of variance (ANOVA), $F_{1,15} = 0.27$, $P = 0.61$). The white lines represent the average and the gray lines the s.e.m. ($n = 16$ sessions collected from 8 rats). NS, not significant. **d**, Choice of high valued side before (Prev) and early and late after delay-only (left) and delay-flavor (right) switches. The rats selected the well where the first reward was delivered after a short delay more often on free-choice trials after learning (two-way ANOVA: chocolate, $F_{2,406} = 152.1$, $P < 0.01$; vanilla, $F_{2,202} = 166.0$, $P < 0.01$). The error bars represent the s.e.m. **e**–**g**, Reaction time (**e**), percentage incorrect choices (**f**) and percentage of trials (**g**) in which rats failed to wait for terminal reward on short (Sh) and long (Lo) forced-choice trials of each flavor. The rats responded faster when the first reward was delivered after a short delay (two-way ANOVA: chocolate, $F_{1,101} = 12.1$, $P < 0.01$; vanilla, $F_{1,101} = 10.8$, $P < 0.01$), made slightly more incorrect responses on short delay trials (two-way ANOVA: chocolate, $F_{1,101} = 5.13$, $P = 0.03$; vanilla, $F_{1,101} = 5.92$, $P = 0.02$) and were equally successful in waiting for the terminal reward on both trial types (two-way ANOVA: chocolate, $F_{1,101} = 0.17$, $P = 0.68$; vanilla, $F_{1,101} = 0.03$, $P = 0.86$). Importantly, there were no main effects or any interactions with block type (delay-only versus delay-flavor) for choice behavior (three-way ANOVA, F values <1.9, P values >0.14, see **d**), reaction time (three-way ANOVA, F values <0.01, P values >0.92 (**e**)) or percentage of incorrect or failed trials (three-way ANOVA, F values <0.19, P values >0.66 (**f**, **g**)). The white lines represent the average and the gray lines the s.e.m. ($n = 102$ sessions collected from 8 rats). **h**, Result of cluster analysis based on the half-time of the spike duration and the ratio comparing the amplitude of the first positive and negative waveform segments ($(n - P)/(n + P)$). Black, reward-responsive putative dopamine (DA) neurons (reward DA, $n = 120$); gray, reward (rew)-nonresponsive putative dopamine neurons (nonrew. DA, $n = 41$); open circle, putative nondopamine neurons (non-DA, $n = 195$). **i**, Violin graphs indicating average amplitude ratio (amp. ratio) and half-time of spike duration of putative dopamine neurons (black) and nondopamine neurons (white). The red lines represent average and the gray lines the s.e.m. ($n = 356$ cells collected from 8 rats).

https://doi.org/10.1038/s41593-023-01310-x

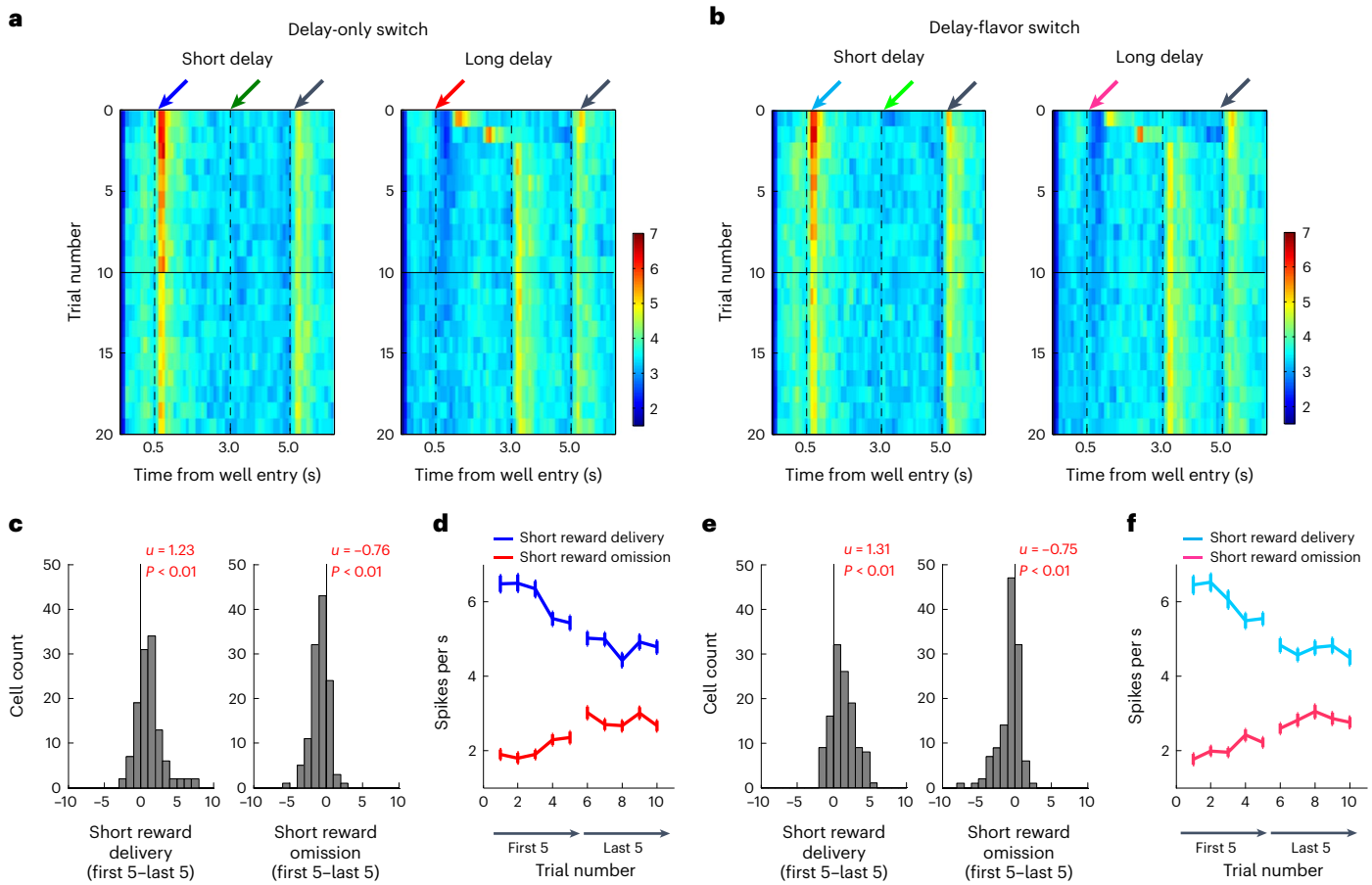


Fig. 2 | Changes in activity of reward-responsive dopamine neurons to shifts in reward timing and flavor. a,**b**, Heat plots representing average activity at the beginning and end of delay-only (**a**) and delay-flavor switch (**b**) blocks. Note that, on trials 1 and 2 of the long-delay blocks, the reward is delayed in steps. The arrows highlight epochs analyzed in the text, including time of delivery of short reward (blue), omission of short reward (red), omission of long reward (green) and delivery of terminal reward (black). **c**,**e**, Distributions of difference scores comparing firing with short reward delivery (left) and omission (right) on the first and last five trials of delay-only (**c**) and delay-flavor switch blocks (**e**). The numbers in each panel indicate the results of two-sided Wilcoxon's signed-rank test (P) and the average difference score (u) ($n = 120$ cells collected from 8 rats). **d**,**f**, Average firing after delivery (blue) and omission of short reward (red) on initial trials and trials at the end of delay-only (**d**) and delay-flavor (**f**) switch blocks. The changes in firing in response to delivery and omission of the short reward were maximal at the beginning of the block and then diminished with learning in both block types (two-way ANOVA, $F > 30.3$, $P < 0.01$). The error bars represent s.e.m. ($n = 120$ cells collected from 8 rats).

the predictive mechanism. We trained rats in an odor-based choice task, in which the timing of one of multiple rewards moved in each block. In some blocks, only the timing of this reward changed; in other blocks, we also changed its identity as it moved in time. As predictions in TDRL models do not include information about the identity of events, this manipulation allowed us to determine whether and how a change in identity is incorporated into the internal models or beliefs reflected in the dopaminergic error signals. Importantly, the occurrence of a different reward early in a trial may be inferred to be a 'bonus' extra reward that does not reset other predictions and thus indicates that the initially expected reward is still to come. Finally, to further test the prediction reset mechanism, we broke the common relationship between the occurrence of a reward and termination of the trial. This was done by adding a second unflavored water reward at the end of every trial at a fixed delay. As we show, this separation revealed that the early appearance of a delayed reward has variable results on prediction-error signaling, which require the dopamine neurons to access multiple, largely independent predictive streams.

Results

Neurons were recorded during performance of an odor-guided choice task in which the rats sampled one of three different odor cues at a central port on each trial, and then had to respond at either a left or a right well to receive two drops of reward (Fig. 1a,b). These three trial types were pseudorandomly ordered and distributed in approximately equal proportions. One odor signaled reward availability in the left well (forced choice left), a second odor signaled reward availability in the right well (forced choice right) and a third odor signaled availability of reward at either well (free choice). In each block of trials, one of two equally preferred flavored milk solutions (chocolate or vanilla, Fig. 1c) was delivered after a short delay (0.5 s) following a correct entry into one well, and the oppositely flavored milk was delivered after a longer delay (1 s on the first trial, 2 s on the second trial and 3 s on subsequent trials) following a correct entry into the other well. In either well, the same amount of water was also delivered at 5 s after correct well entries, serving to terminate trials.

Rats were trained on the task before the start of recording, and, then, during recording, we manipulated the delay to the first reward and its flavor across four transitions between five blocks of trials (Fig. 1b). At the transitions after the first and fourth blocks, both delay and flavor of the first reward were switched, whereas only delay was switched at the other two block transitions. This allowed us to directly compare, for each neuron, its response to changes in reward timing

https://doi.org/10.1038/s41593-023-01310-x

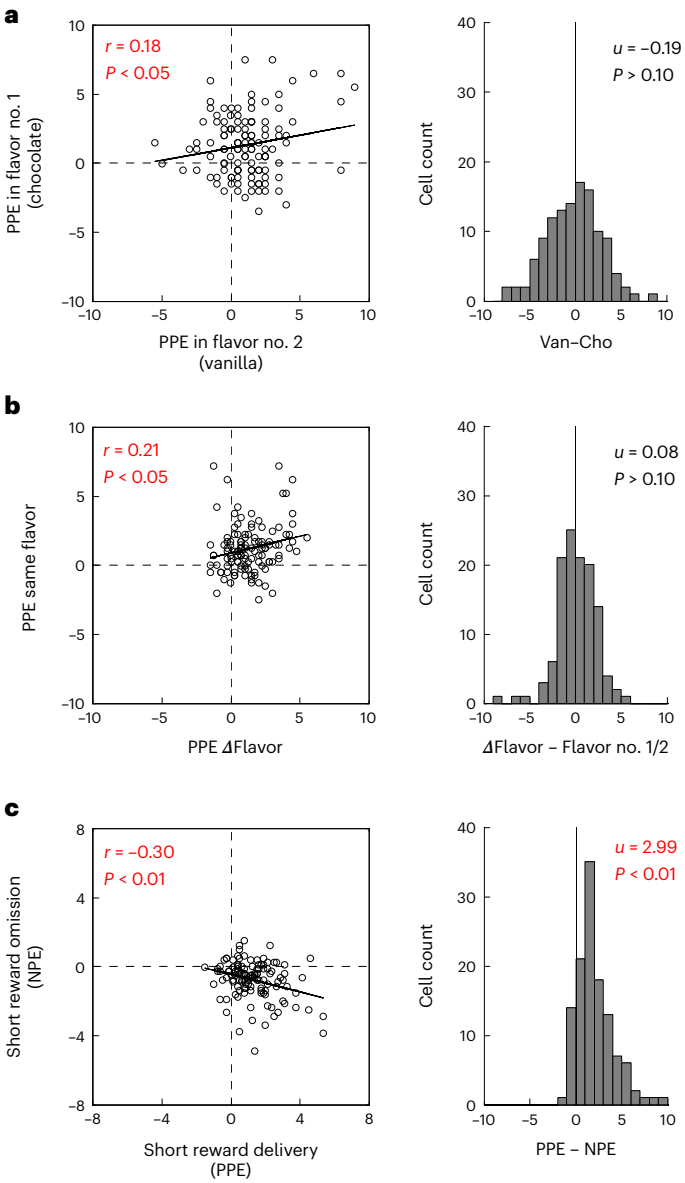


Fig. 3 | Correlations between activity on unexpected reward delivery and omission. a–c, Correlations of difference scores representing changes in firing to unexpected early delivery of chocolate or vanilla in delay-only blocks (**a**), unexpected early delivery of reward in delay-only versus delay-flavor blocks (**b**) and unexpected early delivery or omission of reward in all blocks (**c**). The numbers in each distribution plot indicate the results of two-sided Wilcoxon's signed-rank test (*P*) and the average difference score (*u*) ($n = 120$ cells collected from 8 rats). PPE, positive prediction error. NPE, negative prediction error.

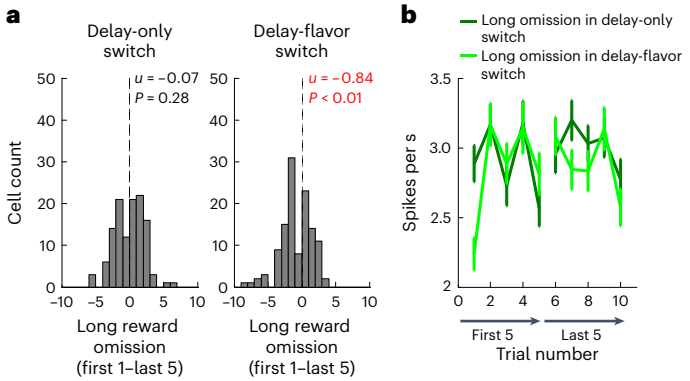


Fig. 4 | Changes in activity of reward-responsive dopamine neurons in response to omission of long reward. a, Distributions of difference scores comparing firing with long reward omission on the initial five versus the last five trials of delay-only (left) and delay-flavor switch blocks (right). The numbers in each panel indicate the results of two-sided Wilcoxon's signed-rank test (*P*) and the average difference score (*u*) ($n = 120$ cells collected from 8 rats). **b**, Average firing after the previously expected but now omitted long reward on the first and last five trials of delay-only (dark green) and delay-flavor switch blocks (light green). Three-way ANOVA (switch type × early/late × trial) showed a significant three-way interaction ($F_{4,476} = 3.16$, $P = 0.01$) and direct comparisons between two types of switches revealed that dopamine neurons fired significantly less on the first omission in the delay-flavor blocks than in the delay-only blocks (two-sided Student's *t*-test, $t = 2.522$, $P = 0.006$). Activity on this specific trial was unlike activity on any of the other trials, as was behavior (licking, Extended Data Fig. 4), and there were no differences in the average firing on any of the other trials ($t < 0.09$, $P > 0.46$). The error bars represent the s.e.m. ($n = 120$ cells collected from 8 rats).

with its response to changes in reward timing combined with reward identity.

The rats responded to both transition types (delay-only and delay-flavor) by preferentially selecting on free-choice trials the well in which the first reward was delivered after a short delay (Fig. 1d). On forced-choice trials, they also responded faster at the well (measured from odor port withdrawal) when the first reward was to be delivered after a short delay (Fig. 1e), although they also made slightly more errors (Fig. 1f). There were no differences in their success at waiting for the terminal reward on correct trials (Fig. 1g) and there were no differences in any of these effects by block type (delay-only versus delay-flavor). See figure captions for supporting statistics.

Dopamine neurons signal errors in response to unexpected delivery or omission of reward

Putative dopamine neurons were identified by means of a cluster analysis based on spike duration and amplitude ratio (Fig. 1h,i); these features were modeled after those used to distinguish dopamine neurons recorded in primates[7], and they have been shown to selectively identify TH^+ (tyrosine hydroxylase-positive) neurons in Long–Evans rats[21]. The cluster analysis identified 161 of 356 neurons recorded in the ventral tegmental area (VTA) as dopaminergic. Of these, 120 increased firing to reward and are analyzed in the main text. As expected, these putative dopamine neurons increased firing whenever a reward was delivered with a timing that was unexpected, as occurred on short delay trials at the start of both delay-only and delay-flavor switch blocks (Fig. 2a,b, dark/light blue arrows). This increase in activity was highest in the first few trials and then declined, consistent with signaling of the prediction error and not the event itself (Fig. 2c–f). A comparison of the errors evoked by early reward according to its flavor and block type showed that, although each difference showed a distribution, they were statistically indistinguishable from zero and the corresponding magnitudes were correlated across the two flavors (Fig. 3a) and block types (Fig. 3b), in accord with the similar valuation and sensory content of the two flavored milk solutions. These same cells also suppressed firing when an expected short reward was delayed (Fig. 2a,b, dark/light red arrows and Fig. 2c–f), which was inversely correlated with the increased firing to unexpected early reward across the population (Fig. 3c). In addition, these cells showed changes in activity to the high and low value cues with learning in each block that correlated with value as indexed by changes in free-choice behavior (Extended Data Fig. 1).

Reset of the dopaminergic error mechanism by early reward is specific rather than global

We next assessed whether these dopamine neurons would lack suppressed firing after the early delivery of a delayed reward, as has been

https://doi.org/10.1038/s41593-023-01310-x

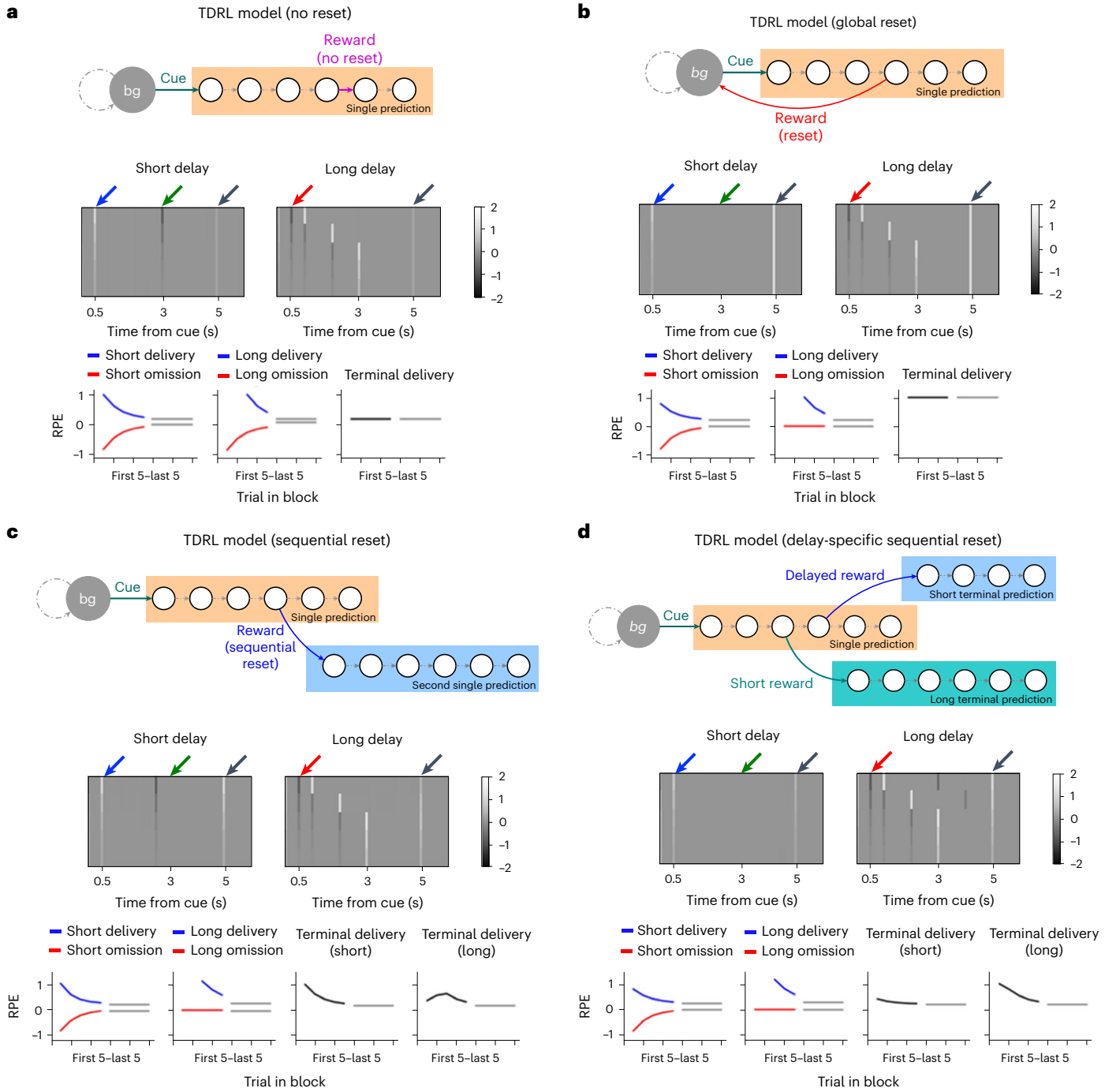


Fig. 5 | Putative reset mechanisms for the classic TDRL model. a, Hidden-state TD model without reset. Trials begin in the background state (*bg*), which acquires no value expectations. On observing the cue, a single sequence of states is initiated. **b**, Hidden-state TD model with global reset. Observing a reward allows a transition to the background state to be inferred. **c**, Hidden-state TD model with sequential reset. Observing a reward allows the initiation of a second sequence of states to be inferred. **d**, Hidden-state TD model with delay-specific sequential reset. Observing a reward up to the short timepoint (0.5 s after the cue) allows the initiation of one secondary sequence of states to be inferred. Observing a reward at a later timepoint allows the initiation of a separate secondary sequence of states to be inferred. Below each model schematic, heat maps show the simulated reward prediction error responses for the corresponding model. Each model predicts the same pattern of reward prediction error (RPE) responses irrespective of the flavor condition of the block. Arrows highlight epochs analyzed in the text, including time of delivery of short reward (blue), omission of short reward (red), omission of long reward (green) and delivery of terminal reward (black). The bottom panels are average RPE at 0.5 s after cue onset (short delivery/omission), 3 s after cue onset (long delivery/omission) and at the time of the terminal reward. Models produced qualitatively similar changes in activity to the cues (Extended Data Fig. 5).

shown in primates[7,15,16]. We found that, when a delayed reward appeared early (without any change in its identity), the dopamine neurons did not suppress firing at the time it had been expected (Fig. 2a, dark green arrow and Fig. 4a, left panel, 4b, darker line). As discussed, this result is inconsistent with the original 'no reset' TDRL models, which predict the suppression of firing at the time of the (now omitted) delayed reward (Fig. 5a, dark green arrow). The absence of this suppression has been explained by incorporating a reset of the reward expectation to zero

https://doi.org/10.1038/s41593-023-01310-x

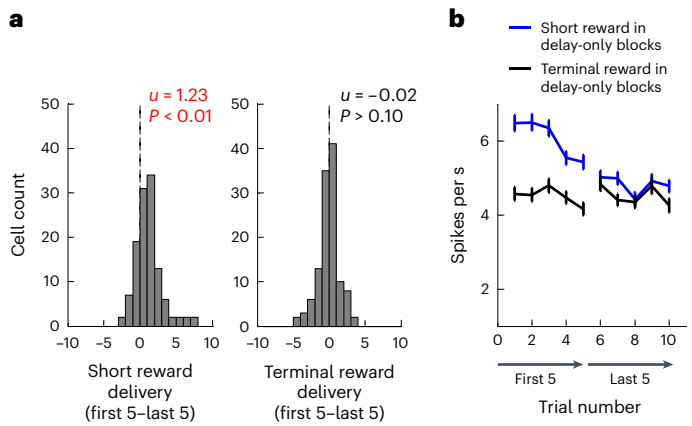


Fig. 6 | Comparison of changes in activity of reward-responsive dopamine neurons to terminal reward and short reward in delay-only blocks. **a**, Distributions of difference scores comparing firing with short (left) and terminal reward on the first and last five trials of short trials in delay-only blocks. The statistics in each panel indicate the results of two-sided Wilcoxon's signed-rank test (P) and the average difference score (u) ($n = 120$ cells collected from 8 rats). **b**, Average firing to short (blue) and terminal (black) reward on the first and last five short trials in delay-only switch blocks. Three-way ANOVA (reward type × early/late × trial) revealed a significant main effect of reward type ($F_{1,118} = 4.87$, $P = 0.03$) and a significant interaction between reward type and early/late ($F_{1,118} = 36.5$, $P < 0.01$). A direct comparison revealed a significant main effect of reward type in early trials (two-way ANOVA, $F_{1,118} = 11.9$, $P < 0.01$), but not in later trials ($F_{1,118} = 0.57$, $P = 0.45$). The error bars represent the s.e.m. ($n = 120$ cells collected from 8 rats).

following the early unexpected appearance of the reward (Fig. 5b, dark green arrow). Notably, however, our data are not fully consistent with this altered model, because such a global reset leaves the subsequent terminal reward unpredicted, resulting in a maximal positive error to this event that does not reduce with experience (Fig. 5b, black arrows). Contrary to this prediction, we observed relatively low activity to the terminal reward when the delayed reward moved early (Fig. 2a,b, black versus blue arrows and Fig. 6a, right panel, 6b, black line).

This global reset model also fails to explain why the same neurons shown to reset in the delay-only blocks (Fig. 2a, dark green arrow) exhibited a significant suppression of firing to the omission of the delayed reward when the flavor was also changed (Fig. 2b, light green arrow and Fig. 4a, right panel, 4b, lighter line). This difference in the neural response to omission of the delayed reward in delay-only versus delay-flavor blocks was observed in both transitions in our five-block design and thus did not depend on the order of training (that is, whether the delay-only or delay-flavor block was first; Extended Data Fig. 2); and it was present even though the subjective value of the two rewards was not different, as per consumption tests and behavior in the task (Fig. 1). Indeed, sorting the blocks according to which flavor was numerically, if nonsignificantly, preferred in this test showed that the effect was present in both situations (Extended Data Fig. 3). Thus, although early delivery of a delayed reward did reset some expectations, this reset appeared to affect only expectations for that specific reward; expectations for the terminal reward remained unchanged, as did expectations for the delayed reward when the identity was not held constant.

For completeness, we also considered whether the delivery of the first reward might reset expectations by initiating a sequential prediction for the terminal reward, following theoretical work in which a reset has been implemented as an inferred transition between hidden states of a task[18]. Such a model did prevent suppression of firing at the time of the (omitted) delayed reward (Fig. 5c, green arrow); however, it also generated a negative error just prior (Fig. 5c, 0.5 s before the green arrow), corresponding to the learned timing of the terminal reward relative to the first reward in this sequential prediction. Furthermore, like the global reset, this model also predicted a large positive error to the terminal reward when it finally did occur on these trials (Fig. 5c, black arrow, short delay blocks). Neither of these features was observed in the dopaminergic response (Fig. 2a,b and Fig. 6).

Finally, we tried a sequential-reset model in which there were two temporally distinct, sequential predictions—one initiated when the first reward occurred at 0.5 s and another when the first reward was delivered later. This model eliminated the suppression of firing at the time of the (omitted) delayed reward without generating an early negative prediction error (Fig. 5d, dark green arrow) and produced subtle effects of reward titration evident on the first two trials of the long-delay blocks (Fig. 5d, trials 1–2 of long blocks); on these trials, there is a negative error of ~2 s after the titrated first reward, reflecting the learned gap between the long and terminal rewards, followed by a positive error to the actual terminal reward when it appears. These errors are weakly present in the actual data (Fig. 2a,b, trials 1–2 of long blocks and Fig. 7c, right panel, 7d, black line), indicating only a subtle influence of these sequential expectations. However, similar to all single-stream models that we considered, this model could not reproduce—even qualitatively—the effect of changes in reward identity on the reset mechanism.

Dopamine neurons access and update multiple predictive streams

One way to achieve a reset that is selective with respect to different, potentially independent outcomes is to allow dopamine neurons to access and update several different, multithreaded predictive streams that reflect not only the timing and value of expected rewards but also their specific features or identity (Fig. 8a,b). In this framework, chains of predictive states are associated with the identity of the outcome they anticipate; even a single cue can support a rich predictive model with multiple threads if associated through experience with different identities of outcomes. Such an architecture would allow an early reward to reset the expectation for the first reward without affecting the expectation for the separate and unique terminal reward in our task (Fig. 8c,d, black arrows). Furthermore, it could also account for a different neural pattern of firing if the moving reward changes its features, as in the delay-flavor blocks of our task. Specifically, a model with access to multiple identity-specific predictive streams should not reset the expectation for a later reward when receiving an early reward of different identity. This is because this early reward may represent an additional reward, not a later reward arriving early. In other words, when the timing alone changes, the dopamine neurons reflect an inference that this represents a movement of the reward; however, when reward identity also changes, they reflect this as the appearance of one reward and the disappearance of another. Consistent with this idea, licking behavior—reflecting an expectation of a reward by the rat at least—was significantly higher during omission of the delayed reward when flavor had changed than when it had not, especially at the start of the block when the maximal neural suppression was observed in the delay-flavor blocks (Extended Data Fig. 4). This pattern of results and the associated modeling suggests that dopamine neurons, or the information streams that they are receiving, consider the sensory features of expected events—in this case, the flavor of a liquid reward—in making the determination of whether a single event has arrived early (and therefore should no longer be expected) or a new event has been introduced. These data, and the relatively stable firing to the terminal reward, indicate that dopamine neurons access a multithreaded predictive model of expected rewards.

Discussion

We recorded dopamine neurons in rats engaged in a new task in which the timing and identity of expected rewards were manipulated. As expected, when a delayed reward was given earlier than expected,

https://doi.org/10.1038/s41593-023-01310-x

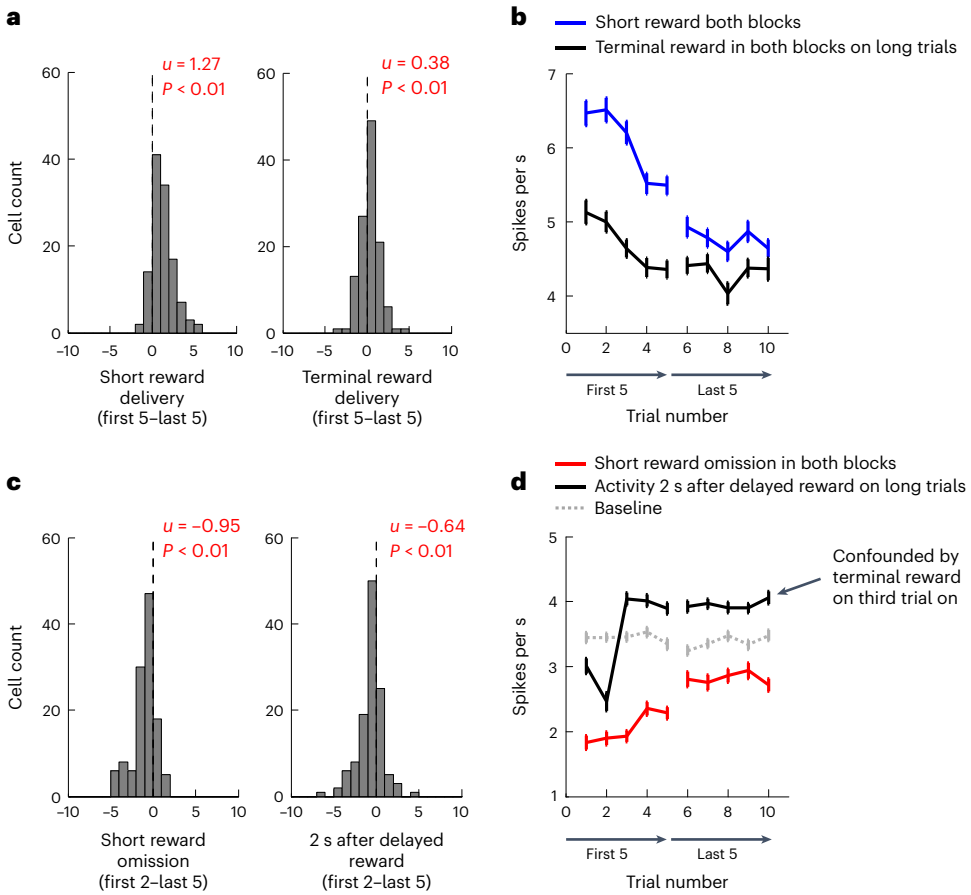


Fig. 7 | Comparison of changes in activity of reward-responsive dopamine neurons to terminal reward and short reward in delay-only blocks. **a**, Distributions of difference scores comparing activity during short reward and terminal reward on the first and last five trials of short trials in both blocks. The statistics in each panel indicate the results of two-sided Wilcoxon's signed-rank test (*P*) and the average difference score (*u*) ($n = 120$ cells collected from 8 rats). **b**, Average firing to short (blue) and terminal (black) reward on the first and last five trials of short trials in both blocks. Four-way ANOVA (reward type × switch type × early/late × trial) revealed significant main effects of reward type ($F_{1,119} = 63.5$, $P < 0.01$), early/late ($F_{1,119} = 88.2$, $P < 0.01$) and trial ($F_{4,476} = 11.6$, $P < 0.01$), and a significant interaction between reward type and early/late ($F_{1,119} = 41.7$, $P < 0.01$). There were no main effects or interactions involving switch type ($F < 1.95$, $P > 0.10$). Direct comparisons revealed that activity was significantly higher for terminal reward in the early and late trials (two-way ANOVA, $F > 12.3$, $P < 0.01$) ($n = 120$ cells collected from 8 rats). **c**, Distributions of difference scores comparing activity during short reward omission (left) and 2 s after delayed reward on the first two and last five trials in both blocks. Note that, as the 2-s period is confounded by delivery of the terminal reward from the third trial on, the last five trials used baseline firing, which is illustrated by the gray dotted line in **d**. The statistics in each panel indicate the results of two-sided Wilcoxon's signed-rank test (*P*) and the average difference score (*u*) ($n = 120$ cells collected from 8 rats). **d**, Average firing to short reward omission (red) and 2 s after delayed reward (black) on the first and last five trials in both blocks. As this 2-s period is confounded by delivery of the terminal reward from the third trial on, the gray dotted line illustrates baseline firing. Four-way ANOVA (reward type × switch type × early/late × trial) revealed significant main effects of reward type ($F_{1,119} = 84.8$, $P < 0.01$), early/late ($F_{1,119} = 50.8$, $P < 0.01$) and trial ($F_{4,476} = 43.9$, $P < 0.01$), and a significant interaction among reward type, early/late and trial ($F_{4,476} = 3.13$, $P = 0.02$). There were no main effects or interactions involving switch type ($F < 1.5$, $P > 0.10$). Direct comparisons revealed that activity was significantly lower for omission of the short reward than in the 2-s period after delayed reward in both the early and the late trials (three-way ANOVA, $F > 50.0$, $P < 0.01$). A direct comparison of activity on the first two trials to short reward omission, the period 2 s after delayed reward and baseline, revealed that each was statistically different from the other two (three-way ANOVA, $F > 23.0$, $P < 0.01$). The error bars represent the s.e.m. ($n = 120$ cells collected from 8 rats).

we found that dopamine neurons responded to the unexpected early reward but did not suppress firing to its later omission, indicating that the later reward was no longer expected after the early arrival. This result replicates findings in primates showing that early reward delivery can reset or nullify expectations for later reward[7,15,16]. However, our findings show that this reset is not general, but applies only to expectations for that specific reward. This specificity was evident in two ways. First, the same neurons showed no changes in firing to a unique and temporally fixed reward that occurred later at the end of each trial and, second, they also suppressed firing on omission of a delayed reward if the early reward was of a different flavor (that is, when a different reward was obtained earlier than expected). As we show, these effects cannot be modeled by current TD algorithms that use, and generate, a single stream of general scalar value predictions; instead they require that dopamine neurons access a multithreaded predictive stream, contributing prediction errors to each thread separately.

Importantly. this is probably just one example of the ability of the biological system—the dopamine neurons—to utilize a more complex and varied state space than is typically envisioned. Although such complexity in the underlying state space is at odds with our current models, it should not be surprising to us as neuroscientists. TDRL models were intended as a simplified heuristic, to illustrate properties of the learning system, but are unlikely to accurately reflect the complexities of the ultimate biology. Although it was sufficient to include one stream of scalar predictions to explain previous data, this should not be taken as a definitive account of the capabilities of the TDRL learning

https://doi.org/10.1038/s41593-023-01310-x

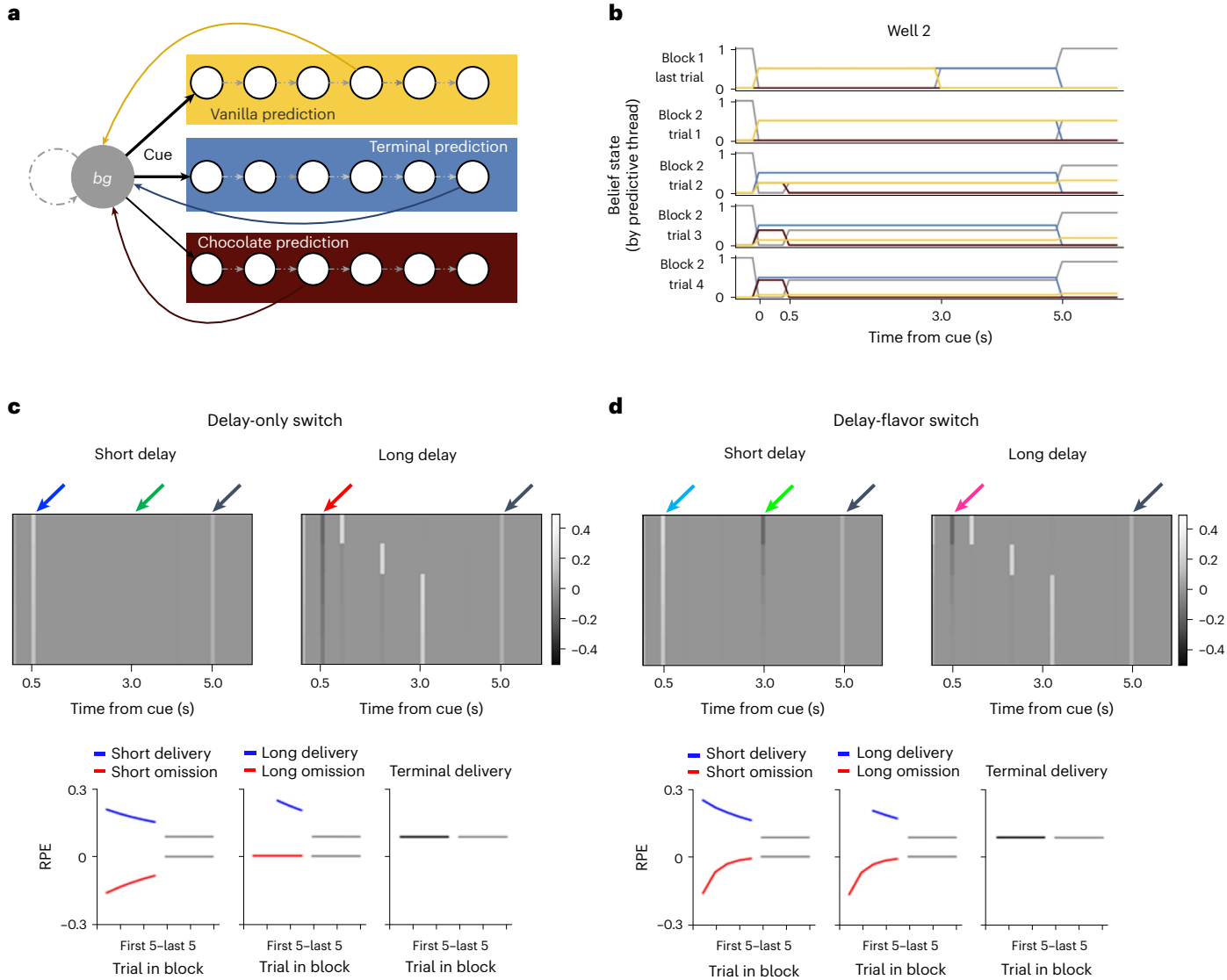


Fig. 8 | Prediction error responses in a TDRL model with multithreaded predictions are sensitive to reward identity. a, Schematic of the hidden-state, multithreaded, prediction TDRL model. Trials begin in the background state (*bg*), which acquires no value expectations. On observing the cue, multiple sequences of states can be initiated. A single sequence, or 'thread', predicts a single outcome identity; on delivery of the correct flavored reward, an identity-specific reset can be inferred. **b**, Simulated belief state for the last trial of block 1 and the first 4 trials of block 2 in well 2. On each trial, the probability of transition from the background state to each of the identity-specific threads is updated according to the rewards that are actually received. **c**,**d**, Heat maps of the simulated reward prediction-error responses after delay-only switch (**c**) and delay-flavor switch (**d**) for the model in which the dopamine neurons have access to a multithreaded predictive model that reflects not only the timing or value of expected rewards but also their features or identity. Arrows highlight epochs analyzed in the text, including time of delivery of short reward (blue), omission of short reward (red), omission of long reward (green) and delivery of terminal reward (black). The bottom panels are average RPE at 0.5 s after cue onset (short delivery/omission), 3 s after cue onset (long delivery/omission) and at the time of the terminal reward. Models produced qualitatively similar changes in activity to the cues (Extended Data Fig. 5).

system without targeted testing of the limitations such a representation implies. Accordingly, the field has recently seen a growing number of correlative and causal reports implicating dopaminergic error signals in situations more complex than these models can easily explain. Two of these are especially relevant to the current results. One involves reports that dopamine neurons signal errors in the prediction of sensory features of expected events in both rats and humans[21–23]. These signals are not salience signals, because they are anti-correlated with the firing of the same neurons on reward omission, which is an obviously salient event[21], and the sensory content of the prediction error can be decoded from their pattern[22]. Instead, such activity appears to reflect a sensory prediction error, an idea supported by causal data showing that dopamine transients are involved in learning in response to unexpected changes in the sensory content of both rewarding[24,25] and neutral events[26]. The current data add to these previous results, confirming that the predictive framework underlying dopaminergic prediction-error firing incorporates information about the sensory features of expected rewards, while further showing that predictions about rewards of different identities can be maintained independently. Such a multithreaded signal is consistent with the need to track multiple independent features of our environments. It is interesting that having independent predictions of rewards of different flavors would give rise to the sensory prediction errors observed in these previous studies[21–23].

A second area in which dopamine has been implicated that is relevant to the current findings is its role in tracking latent causes or so-called belief states[20,27,28]. The current data support this proposal because the determination of whether to recognize the early reward in our task as either relocated from later or new is entirely based on an internal belief of the subject. It is akin to the visual perception trick in which a dot moves behind a piece of paper, only to emerge shortly thereafter on the other side. Is it the same dot or a new one? Visual processing areas integrate variables regarding speed, size, color and our understanding of how the world works to infer the answer to this question. This same sort of inference process is at work in the current experiment, and evident in the subjects' behavior and the recorded dopamine neurons' firing patterns; dopamine neuron activity in the delay-only and delay-flavor blocks reflects the impact of this internal belief state, which is probably dependent on input from a subset of areas impacting on the VTA, particularly the prefrontal regions key to inference and creating cognitive maps involving hidden states[29] and known to impact dopamine neuron signaling[30–33].

Consistent with this emphasis on belief state, it is worth noting that we do not intend to imply that the flavor of an expected reward or even its external identity is the only variable that might be used to distinguish predictive streams. Flavor is just the feature we used here; any other feature might contribute to determining that an expected event is unique and unrelated to other events and thus deserving of its own predictive stream, including features relevant to value such as size, number, current need and even the general context of the task. For instance, if some trials involved three rewards, an early reward may be treated as an extra one, rather than the delayed reward arriving early. Conversely, it is possible that dopamine neurons in the current design would behave very differently if the subject were extremely deprived, which might promote generalization across the two flavors. Indeed, our data probably already reflect some degree of generalization; the suppression is clearly evident on the first trial but not thereafter, suggesting partial but not complete separation of the predictive streams in our particular task. Similarly, we do not require the multiple predictions evident in our data to operate entirely in parallel. Although not modeled here, one can imagine designs in which sequential predictions between the different rewards, as we have modeled, might be nested within or coexist with parallel predictive streams. Indeed, although we did not do it here, the principle of multithreaded predictions can be easily extended to more sophisticated task representations such as semi-Markov state representations[19,34], which will then be able to account for both the temporal and the sequential dependence of reward predictions, as well as the complex outcome specificity of these predictions that we demonstrate here. These unanswered questions notwithstanding, our findings show concretely that dopamine neurons parse much more detailed, multilayered and complex predictive input than typically envisioned. This has implications not only for the various sources of afferent input, which must be less interchangeable and contain higher-resolution information than currently thought, but also for the output, which is likely to be distributed and multidimensional, providing much more information to downstream areas regarding prediction errors than currently assumed.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-023-01310-x.

References

1. Schultz, W. Dopamine reward prediction-error signalling: a two-component response. *Nat. Rev. Neurosci.* **17**, 183–195 (2016).
2. Keiflin, R. & Janak, P. H. Dopamine prediction errors in reward learning and addiction: from theory to neural circuitry. *Neuron* **88**, 247–263 (2015).
3. Glimcher, P. W. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proc. Natl Acad. Sci. USA* **108**, 15647–15654 (2011).
4. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate for prediction and reward. *Science* **275**, 1593–1599 (1997).
5. Watabe-Uchida, M., Eshel, N. & Uchida, N. Neural circuitry of reward prediction error. *Annu. Rev. Neurosci.* **40**, 373–394 (2017).
6. Mirenowicz, J. & Schultz, W. Importance of unpredictability for reward responses in primate dopamine neurons. *J. Neurophysiol.* **72**, 1024–1027 (1994).
7. Hollerman, J. R. & Schultz, W. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.* **1**, 304–309 (1998).
8. Waelti, P., Dickinson, A. & Schultz, W. Dopamine responses comply with basic assumptions of formal learning theory. *Nature* **412**, 43–48 (2001).
9. Tobler, P. N., Dickinson, A. & Schultz, W. Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *J. Neurosci.* **23**, 10402–10410 (2003).
10. Lak, A., Stauffer, W. R. & Schultz, W. Dopamine prediction error responses integrate subjective value from different reward dimensions. *Proc. Natl Acad. Sci. USA* **111**, 2342–2348 (2014).
11. Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B. & Uchida, N. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* **482**, 85–88 (2012).
12. Eshel, N. et al. Arithmetic and local circuitry underlying dopamine prediction errors. *Nature* **525**, 243–246 (2015).
13. Pan, W.-X., Schmidt, R., Wickens, J. R. & Hyland, B. I. Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. *J. Neurosci.* **25**, 6235–6242 (2005).
14. Kim, H. R. et al. A unified framework for dopamine signals across timescales. *Cell* **183**, 1600–1616 (2020).
15. Fiorillo, C. D., Newsome, W. T. & Schultz, W. The temporal precision of reward prediction in dopamine neurons. *Nat. Neurosci.* **11**, 966–973 (2008).
16. Kobayashi, K. & Schultz, W. Influence of reward delays on responses of dopamine neurons. *J. Neurosci.* **28**, 7837–7846 (2008).
17. Suri, R. E. & Schultz, W. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience* **91**, 871–890 (1999).
18. Daw, N., Courville, A. C. & Touretzky, D. S. Representation and timing in theories of the dopamine system. *Neural Comput.* **18**, 1637–1677 (2006).
19. Takahashi, Y. K., Langdon, A. J., Niv, Y. & Schoenbaum, G. Temporal specificity of reward prediction errors signaled by putative dopamine neurons in rat VTA depends on ventral striatum. *Neuron* **91**, 182–193 (2016).
20. Starkweather, C. K., Babayan, B. M., Uchida, N. & Gershman, S. J. Dopamine reward prediction errors reflect hidden-state inference across time. *Nat. Neurosci.* **20**, 581–589 (2017).
21. Takahashi, Y. K. et al. Dopamine neurons respond to errors in the prediction of sensory features of expected rewards. *Neuron* **95**, 1395–1405 (2017).
22. Stalnaker, T. A. et al. Dopamine neuron ensembles signal the content of sensory prediction errors. *eLife* **8**, e49315 (2019).
23. Howard, J. D. & Kahnt, T. Identity prediction errors in the human midbrain update reward-identity expectations in the orbitofrontal cortex. *Nat. Commun.* **9**, 1–11 (2018).

24. Chang, C. Y., Gardner, M., Di Tillio, M. G. & Schoenbaum, G. Optogenetic blockade of dopamine transients prevents learning induced by changes in reward features. *Curr. Biol.* **27**, 3480–3486 (2017).
25. Keiflin, R., Pribut, H. J., Shah, N. B. & Janak, P. H. Ventral tegmental dopamine neurons participate in reward identity predictions. *Curr. Biol.* **29**, 92–103 (2019).
26. Sharpe, M. J. et al. Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nat. Neurosci.* **20**, 735–742 (2017).
27. Lak, A., Nomoto, K., Keramati, M., Sakagami, M. & Kepecs, A. Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. *Curr. Biol.* **27**, 821–832 (2017).
28. Starkweather, C. K. & Uchida, N. Dopamine signals as temporal difference errors: recent advances. *Curr. Opin. Neurobiol.* **67**, 95–105 (2021).
29. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267–279 (2014).
30. Starkweather, C. K., Gershman, S. J. & Uchida, N. The medial prefrontal cortex shapes dopamine reward prediction errors under state uncertainty. *Neuron* **98**, 616–629 (2018).
31. Jo, Y. S. & Mizumori, S. J. Prefrontal regulation of neuronal activity in the ventral tegmental area. *Cereb. Cortex* **26**, 4057–4068 (2016).
32. Jo, Y. S., Lee, J. & Mizumori, S. J. Effects of prefrontal cortical inactivation on neural activity in the ventral tegmental area. *J. Neurosci.* **33**, 8159–8171 (2013).
33. Takahashi, Y. K. et al. Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. *Nat. Neurosci.* **14**, 1590–1597 (2011).
34. Langdon, A. J., Sharpe, M. J., Schoenbaum, G. & Niv, Y. Model-based predictions for dopamine. *Curr. Opin. Neurobiol.* **49**, 1–7 (2017).

Methods

These experiments received ethical approval from the National Institute on Drug Abuse Animal Care and Use Committee under animal study protocols 18-CNRB-108 and 20-CNRB-108.

Subjects

Eight male Long–Evans rats (Charles River Labs), aged approximately 3 months at the start of the experiment, were used in the present study.

Stereotaxic surgery

All surgical procedures adhered to guidelines for aseptic technique. For electrode implantation, a drivable bundle of eight 25-µm diameter FeNiCr wires (Stablohm 675, California Fine Wire) was chronically implanted dorsal to the VTA in the left or right hemisphere at 5.3 mm posterior to bregma, 0.7 mm laterally and 7.5 mm ventral to the brain surface at an angle of 5° toward the midline from vertical. Wires were cut with surgical scissors to extend ~2.0 mm beyond the cannula and electroplated with platinum (H_2PtCl_6, Sigma-Aldrich) to an impedance of 500–900 kΩ. Cephalexin (15 mg kg^{-1} orally) was administered twice daily for 2 weeks postoperatively.

Histology

All rats were perfused with phosphate-buffered saline followed by 4% paraformaldehyde (Santa Cruz Biotechnology Inc.). Brains that received only electrode implantation were cut in 40-µm sections and stained with thionin.

Odor-guided choice task

Recording was conducted in aluminum chambers. A central odor port was located in one wall above two fluid wells. Two lights were located above the panel. The odor port was connected to an air flow dilution olfactometer to allow the rapid delivery of olfactory cues. Odors were chosen from compounds obtained from International Flavors and Fragrances. Trials were signaled by illumination of the panel lights inside the box. When these lights were on, nosepoke into the odor port resulted in delivery of the odor cue to a small hemicylinder located behind this opening. One of three different odors was delivered to the port on each trial, in a pseudorandom order. At odor offset, the rat had 3 s to make a response at one of the two fluid wells. One odor instructed the rat to go to the left to get a reward, a second odor instructed the rat to go to the right to get a reward and a third odor indicated that the rat could obtain a reward at either well. Odors were presented in a pseudorandom sequence such that the free-choice odor was presented on 7 of 20 trials and the left/right odors were presented in equal numbers. In addition, the same odor could be presented on no more than three consecutive trials. Once the rats were shaped to perform this basic task, we introduced blocks in which we manipulated the delay preceding reward delivery (Fig. 1b). For recording, one well was randomly designated as short and the other as long at the start of the session (Fig. 1b). On short trials, a drop of chocolate- or vanilla-flavored milk was delivered at 0.5 s and a drop of water was delivered at 5.0 s after the rats entered the fluid wells. On long trials, a drop of the other-flavored milk was delivered at 1.0 s on the first trial, 2.0 s on the second trial and 3.0 s on the third trial and thereafter. In addition, a drop of water was also delivered at 5.0 s on all trials; this terminal reward ensured that the rats would remain in the well during the delays on both the short and the long trial types. Once rats exhibited biased choice behavior in the initial block, we switched the delays to the flavored reward in the two wells, across four transitions. At the first and fourth transitions, we also switched the flavor of the two rewards when we switched the delays. All blocks were 30- to 50-trials long; block switches were triggered when rats chose the short reward side on at least seven of the last ten free-choice trials.

Single-unit recording

Wires were screened for activity daily; if no activity was detected, the rat was removed and the electrode assembly was advanced 40 or 80 µm. Otherwise, active wires were selected to be recorded, a session was conducted and the electrode was advanced at the end of the session. Neural activity was recorded using Plexon Multichannel Acquisition Processor systems. Signals from the electrode wires were amplified 20× by an op-amp headstage (Plexon Inc., catalog no. HST/8o50-G20-GR), located on the electrode array. Immediately outside the training chamber, the signals were passed through a differential preamplifier (Plexon Inc., catalog no. PBX2/16sp-r-G50/16fp-G50), where the single-unit signals were amplified 50× and filtered at 150–9,000 Hz. The single-unit signals were then sent to the Multichannel Acquisition Processor box, where they were further filtered at 250–8,000 Hz, digitized at 40 kHz and amplified at 1–32×. Waveforms (>2.5:1 signal:noise) were extracted from active channels and recorded to disk by an associated workstation.

Data analysis

Units were sorted using Offline Sorter software from Plexon Inc. Sorted files were then processed and analyzed in NeuroExplorer and MATLAB. Dopamine neurons were identified via a waveform analysis. Briefly cluster analysis was performed based on the half-time of the spike duration and the ratio comparing the amplitude of the first positive and negative waveform segments. The center and variance of each cluster were computed without data from the neuron of interest, and then that neuron was assigned to a cluster if it was within 3 s.d. of the cluster's center. Neurons that met this criterion for more than one cluster were not classified. This process was repeated for each neuron. Putative dopamine neurons that showed an increase in firing to reward compared with baseline (400 ms before reward) were further classified as reward responsive (Student's t-test, $P < 0.05$). To analyze neural activity to reward, we examined firing rate in the 400 ms beginning 100 ms after reward delivery.

Computational modeling

We modeled dopaminergic prediction-error signals during the task using variations of the original temporal difference (TD) algorithm[35,36], applied in a partially observable environment[18,20,37]. In TD learning, reward predictions take the form of a value expectation V, which is the expected cumulative discounted future reward at each timepoint t:

$$V(t) = E\left[\sum_{\tau=t}^{\infty} \gamma^{\tau-t} R(\tau)\right],$$

where $0 \leq \gamma \leq 1$ exponentially discounts reward R delivered at τ and $E[\cdot]$ denotes the expectation (that is, the average).

Rather than model timepoints during a trial using the complete serial-compound (CSC) representation[4,38], we assume that each trial is represented as a Markovian sequence of hidden (that is, latent) discrete states $s \in \{1, \ldots, N\}$, in which the current state must be inferred using knowledge of the transition structure between hidden states (described by the transition matrix $T_{s_t \to s_{t+1}} = P(s_{t+1}|s_t)$) and the relationship between the possible observations that occur during the task and state transitions during a trial (described by the observation matrices $O_{s_t \to s_{t+1}}[o_{t+1}] = P(o_{t+1}|s_{t+1}, s_t)$). This partially observable setting allows us the flexibility to model different criteria for inferring a reset of reward predictions contingent on the observations made during a trial, along with different assumptions about the latent, possibly distributed, structure of reward predictions in the task. Thus, rather than learning a value for each timepoint, values are learned for each element of the belief state, which is the probability distribution over hidden state

occupancy at each timepoint. Following Bayes' rule, the belief state evolves from t to $t+1$ according to:

$$b_s(t+1) = P(s_{t+1}|o_{t+1}, s_t) \propto \sum_{s_t} P(o_{t+1}|s_{t+1}, s_t) P(s_{t+1}|s_t) P(s_t)$$

where the proportionality is resolved by normalizing the probability of all possible states s_{t+1} to sum to 1.

In each model, we assume a single 'background' state (denoted *bg*) that does not accrue any value (that is, cannot be associated with a reward prediction), whereas all other states are grouped into 'threads'—chains of states (denoted by c; see below) through which the model progresses one timepoint after another. This progression can be interrupted by inferring a reset—a transition to the zero-value background state rather than continued progression within a predictive thread—contingent on observation of certain events (specific details for each model are given below). We model only reward expectation during the task and not the choice, by representing entry to each reward well as a cue. Throughout, we assume the space of possible observations is fixed, with $o \in \{\text{empty (that is, no event)}, \text{cue}, R_{\text{vanilla}}, R_{\text{chocolate}}, R_{\text{terminal}}\}$.

Single-thread classic TDRL model without reset. In our single-thread TDRL model, the estimated value at time t is the linear combination of learned predictive weights over the belief state at the current timepoint:

$$V(t) = \sum_s w_s(t)\, b_s(t).$$

Value estimates for the current and next timestep are used to compute a prediction error:

$$\delta(t) = R_{t+1} + \gamma V(t+1) - V(t),$$

which is used to update the predictive weights via a TD(λ) learning rule:

$$w_s(t+1) = w_s(t) + \eta e_s \delta(t),$$

with learning rate $0 \le \eta \le 1$. We decay all learned weights according to $w_s \leftarrow (1-\kappa)\, w_s$ between successive trials with decay constant $0 \le \kappa \le 1$ (where 1 is complete decay of learned predictive weights to 0 over a single timestep and 0 is no decay). Although decay is not typically included in the classic TD algorithm, we include it here to account for the residual phasic dopaminergic response to predicted rewards that is apparent in the empirical results. As mentioned above, the predictive weight for the background state is fixed at 0. This TD(λ) learning rule uses eligibility traces, e_s, to allow the update of predictive weights for states that have been occupied previous to the current timepoint in a given trial. Eligibility traces evolve according to:

$$e_s(t) = \gamma\lambda e_s(t-1) + b_s(t),$$

with the constraint that $0 \le e_s \le 1$. This means the eligibility trace for a state is set to the current belief in occupancy of that state (a maximum of 1), and then decays with time constant λ over future timepoints during a trial.

Importantly, and following the complete serial-compound stimulus representation commonly used in TDRL models[39], the transition and observation probabilities of this model do not include a reset of the single chain of predictive states once this chain has been initiated by observing the cue. State transitions in this model are $T_{bg\to bg} = 0.5$ (for the transition from the background state to itself), $T_{bg\to c_0} = 0.5$ (initiating the single thread of states corresponding to a trial), $T_{c_j\to c_{j+1}} = 1$ (continuing the thread) and $T_{c_N\to bg} = 1$ (thread termination back to the background state, for the intertrial interval), where c_j denotes the ordered states within a predictive thread ($j = 0, \ldots, N$ states). All other transition probabilities are 0, meaning that all trials have the same structure traversing the single thread of consecutive states. The observation matrices contain the probability of each unique observation being emitted during the transition from $s_t \to s_{t+1}$; only observation probabilities that correspond to permissible transitions are nonzero: $O_{bg\to bg}[\{\text{empty}, R_{\text{vanilla}}, R_{\text{chocolate}}, R_{\text{terminal}}\}] = 0.25$, $O_{bg\to c_0}[\text{cue}] = 1$, $O_{c_j\to c_{j+1}}[\{\text{empty}, R_{\text{vanilla}}, R_{\text{chocolate}}, R_{\text{terminal}}\}] = 0.25$ and $O_{c_N\to bg}[\text{empty}] = 1$.

Single-thread classic TDRL model with reset. Within the single-thread state representation, we model a reset of reward predictions as an inferred transition to the zero-value background state of the task. This inference process is controlled by the combination of state transition probabilities contained in T and the transition-specific observation probabilities in O. The within-predictive chain transition structure is modified from the previous model to allow for the equal possibility of either continuation or reset from any state c_j: $T_{c_j\to bg} = 0.5$, $T_{c_j\to c_{j+1}} = 0.5$. Whether a reset is actually inferred at any timepoint depends on the current observation; in this model, we apply a reset on observing any one of the three possible rewards by allowing an observation of a reward to be possible only with a transition to the background state: $O_{c_j\to bg}[\{R_{\text{vanilla}}, R_{\text{chocolate}}, R_{\text{terminal}}\}] = 0.33$, $O_{c_j\to c_{j+1}}[\{R_{\text{vanilla}}, R_{\text{chocolate}}, R_{\text{terminal}}\}] = 0$ and $O_{c_j\to c_{j+1}}[\text{empty}] = 1$, that is, continued progression within the single thread of predictive states once it has been initiated is possible only with an accompanying empty observation.

An inferred transition to the background state also resets eligibility traces e_s for all states within the single chain:

$$e_s = e_s \cdot \sum_s b_s(t),$$

for all $s \in \{c_0, \ldots, c_N\}$. This means that the eligibility of any state for update is limited by the maximum state occupancy in that thread; once a (complete) reset has been inferred, this term is zero, effectively wiping out the eligibility of within-thread states for future updates of their predictive weights.

Single-thread classic TDRL model with sequential reset. We also consider a model in which a reset involves the initiation of a new chain of states on delivery of the first reward, rather than a transition to the nonpredictive background state. This sequential-reset model has only a single thread active at any point during the trial, and the second thread is initiated on observing the first reward (of either flavor). The transition structure for this model is amended from the previous single-chain models to allow for the equal probability of continuation of the first thread (A) or initiation of the second thread (B): $T_{A_j\to B_1} = 0.5$, $T_{A_j\to A_{j+1}} = 0.5$, where B_1 is the first state of the second thread. Probability of continuation of the second thread is split with the probability of transition to the background like the previous reset model: $T_{B_j\to bg} = 0.5$, $T_{B_j\to B_{j+1}} = 0.5$. Here, the null observation is uniquely mapped to the continuation of A or B, whereas observation of a reward is associated with transition from A to B and B to background.

Single-thread classic TDRL model with delay-specific sequential reset. This model is a modification of the single-thread sequential-reset model above, in which one of two distinct threads is initiated on receiving the first reward, depending on the delay to that first outcome. Transition structure for this model allows for the equal probability of continuation of the first thread (A) or initiation of the second thread (B): $T_{A_j\to B_1} = 0.5$, $T_{A_j\to A_{j+1}} = 0.5$, for all states j up to and including $t = 0.5$ s, and continuation or initiation of a third thread C for all states j later than this timepoint: $T_{A_j\to C_1} = 0.5$, $T_{A_j\to A_{j+1}} = 0.5$.

Note that T and O for these single-thread TDRL models are fixed and have been constructed to result in low state uncertainty; a single discrete state is occupied at each timepoint to make these models equivalent to TD models developed with the CSC representation, while

allowing for an inferred reset of the hidden state of the task based on the sequence of observations during a trial.

Multithread TDRL model. To account for the identity specificity of the dopaminergic prediction-error responses, we modeled multiple predictive threads of task states, $c_j[A]$, where j indexes the ordered states within a thread (as above) and A indexes the different threads. We introduce a new TD(λ) algorithm in which value expectations are specific to the expected outcome identities:

$$V_A(t) = \sum_s M_{A,s} w_s(t)\, b_s(t),$$

where $A \in$ {vanilla, chocolate, terminal} for this task. The matrix $M_{A,s}$ controls the mapping between states and the identity-specific value expectations to which they contribute; for simplicity, we assume that a thread of reward-predictive states does not generalize across different outcome expectations, requiring all states within a single predictive thread A to map to a single outcome identity. This means that we constrain $M_{A,s}$ to be 1 for all states within a predictive thread A and 0 otherwise. Throughout the main text we thus name each predictive thread by the outcome identity to which it uniquely maps. (Note that predictive weights in this case need only be indexed by state s.)

Prediction errors in this model are thus also vectorized, with one element per outcome identity:

$$\delta_A(t) = R_A(t+1) + \gamma V_A(t+1) - V_A(t).$$

This vector prediction-error signal is used to update the predictive weights of each state:

$$w_s(t+1) = w_s(t) + \eta e_s\left(U_{s,A}\delta_A(t)\right),$$

where $U_{s,A}$ controls the mapping between prediction errors in each channel and states. In the present study, we take $U_{s,A} = M^T_{A,s}$ (that is, the inverse of the matrix $M_{A,s}$) such that the identity-specific prediction-error maps back to the corresponding thread of predictive states. Eligibility traces for each state follow the same dynamics as previously described and weights decay between successive trials by $w_s \leftarrow (1-\kappa)\, w_s$ for all states s that had nonzero eligibility during the trial. Note that this TD learning rule reduces to the single thread update for $A \in \{1\}$ and $M_{A,s} = 1$ for all s. Here, noting the correlated sensitivity of individual units to rewards of different identities, we assume that the identity-specific vector prediction-error signal is distributed across the population of dopamine neurons and model the average population response as the average over threads for comparison to the data.

The transition probabilities for the multithread TDRL model control which predictive threads are initiated on observation of the cue. The transition matrix is similar to that for the single-thread TDRL model with reset, with $T_{c_j[A]\to c_{j+1}[A]} = 0.5$ and $T_{c_j[A]\to bg} = 0.5$ within each thread, and all transitions between threads set to 0, that is, $T_{c_j[A]\to c_k[B]} = 0$. Unlike the previous models, which have fixed transition probabilities, in the present study we also update the probability of initiating a specific thread of predictive states based on the observations made within each trial. At the completion of a trial, transition probabilities to initiate each predictive thread are thus learned as follows:

$$T_{bg\to c_0[A]} = T_{bg\to c_0[A]} + \eta_T\left[1 - T_{bg\to c_0[A]}\right]$$

if the outcome corresponding to A was delivered on that trial, and

$$T_{bg\to c_0[A]} = T_{bg\to c_0[A]} + \eta_T\left[0 - T_{bg\to c_0[A]}\right]$$

otherwise[40].

Observation probabilities for transitions to and from the background state in this model are similar to the single-thread TDRL model with reset, with $O_{bg\to bg}[\text{empty}] = 0.5$, $O_{bg\to bg}[\{R_{\text{vanilla}}, R_{\text{chocolate}}, R_{\text{terminal}}\}] = 0.167$ and $O_{bg\to c_0[A,B,C]}[\text{cue}] = 0.33$. To determine which specific observations trigger a reset of each predictive thread by allowing a transition to the background state to be inferred, we set for each predictive thread a nonzero probability of observing the specific outcome corresponding to the channel identity on transition to the background state: $O_{c_j[A]\to bg}[\{R_A\}] = 1$. In addition, $O_{c_j[A]\to c_{j+1}[A]}[\{\text{empty}\}] = 0.5$, and $O_{c_j[A]\to c_{j+1}[A]}[\{R_B, R_C\}] = 0.25$, where B and C are the outcome types that do not match the corresponding channel identity of predictive thread A.

Model simulations. For simplicity, and to explain the recording results presented in this manuscript, we model only the processes of reward prediction and learning, ignoring the (separate) mechanism of action selection. For this, we used a simplified representation of the task in which entry into a reward well is modeled as the onset of a cue (specific to that well) and reward delivery follows the same temporal and identity contingencies as in the empirical task, with a fixed block length of 50 trials. We simulated each trial using a timestep of 0.1 s for a total duration of 9 s. Parameters for the single-thread learning models were $\eta = 0.4$, $\gamma = 0.95$, $\lambda = 1$ and $\kappa = 0.1$. Parameters for the multithread TDRL model were the same, with the addition of $\eta_T = 0.5$ for learning the transition probabilities between the cue and the start of each thread.

Statistics and reproducibility

This section has been added to comply with journal policies. The study was designed to allow each session to contain relevant manipulations to address the a priori hypothesis motivating the study, thus most of our analyses were within sessions and subjects/neurons. Within this framework, the presentation of cues and trial types, and precise timing of the switches and the organization of the initial trial block were chosen pseudorandomly, by rules in the computer program running the task or by an experimenter when outside the session. The experimenter was not blind to the conditions of the experiment. No statistical method was used to predetermine sample size or specific tests applied; rather, sample sizes were chosen, based on previous similar work, as those required to demonstrate robust prediction-error correlates, and statistical tests were chosen based on their appropriateness for the design and to maintain consistency with previous studies[19,21,33]. No animals or data points were excluded from the analyses. Data distribution was assumed to be normal, but this was not formally tested.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The dataset and all scripts used in the present study are available at https://github.com/YouGTakahashi/ultra_delay_analysis for the unit analyses and at https://github.com/ajlangdon/multithreadTD for the modeling.

Code availability

The dataset and all scripts used in the present study are available at https://github.com/YouGTakahashi/ultra_delay_analysis for the unit analyses and at https://github.com/ajlangdon/multithreadTD for the modeling.

References

35. Sutton, R. S. Learning to predict by the method of temporal difference. *Mach. Learn.* **3**, 9–44 (1988).
36. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An introduction* (MIT Press, 1998).

37. Kaelbling, L. P., Littman, M. L. & Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artif. Intelligence* **101**, 99–134 (1998).
38. Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive hebbian learning. *J. Neurosci.* **16**, 1936–1947 (1996).
39. Ludvig, E. A., Sutton, R. S. & Kehoe, E. J. Evaluating the TD model of classical conditioning. *Learn. Behav.* **40**, 305–319 (2012).
40. Glascher, J., Daw, N., Dayan, P. & O'Doherty, J. P. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).

Acknowledgements

This work was supported by grant no. Z1A-DA000587 (to G.S.) at the Intramural Research Program of the National Institute on Drug Abuse, the Intramural Research Program of the National Institute of Mental Health (to A.J.L.) and by grant no. U01DA050647 (to A.J.L.) from the National Institute on Drug Abuse. The opinions expressed in this article are the authors' own and do not reflect the view of the National Institutes of Health/Department of Health and Human Services.

Author contributions

Y.K.T., T.A.S. and G.S. designed the experiments. Y.K.T. and L.E.M. conducted the behavioral training and single-unit recording. S.K.H. and A.J.L. conducted the modeling. Y.K.T., A.J.L. and G.S. interpreted the data and wrote the manuscript with input from the other authors.

Competing interests

The authors declare no competing interests.

Additional information

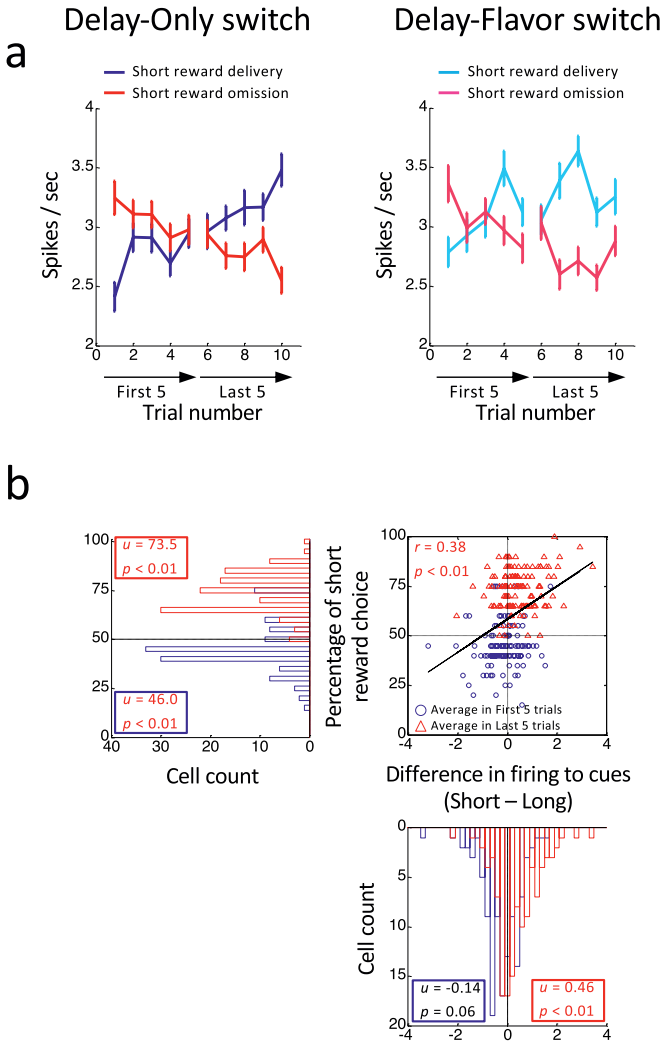
Extended data is available for this paper at https://doi.org/10.1038/s41593-023-01310-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-023-01310-x.

Correspondence and requests for materials should be addressed to Yuji K. Takahashi, Angela J. Langdon or Geoffrey Schoenbaum.

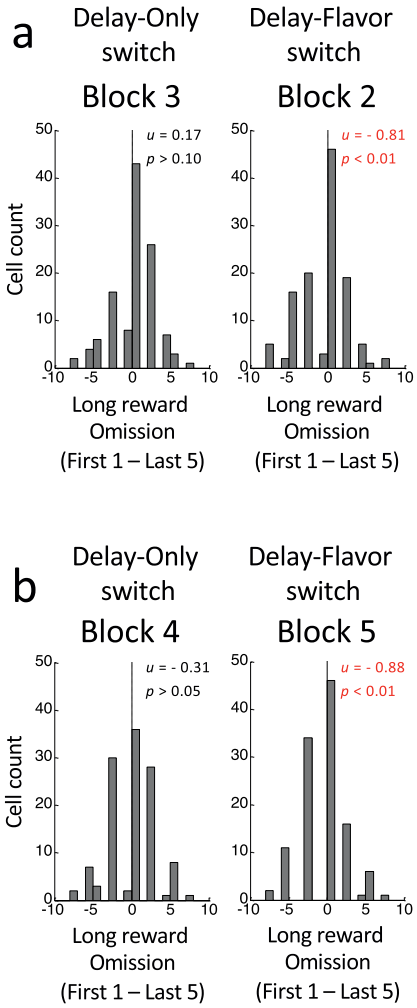
Peer review information *Nature Neuroscience* thanks Okihide Hikosaka, Kevin Miller and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41593-023-01310-x



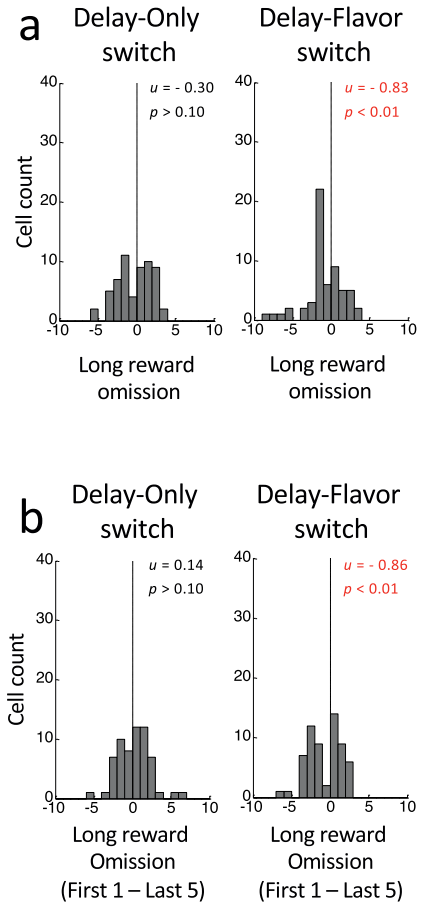
Extended Data Fig. 1 | Activity to the odor cues changes with value and free choice behavior across trial blocks and does so similarly for the two block types. (**a**) Average firing in the reward-responsive dopamine neurons during presentation of the high and low value cues. Average firing is plotted for the first and last 5 trials of the Delay-Only and Delay-Flavor blocks. Activity increased to the high value cue, paired with the early reward, across each block (Three-way ANOVA, Trial x value, $F_{9,1071} = 9.46$, $p < 0.01$), and there was no difference in the pattern across switch types ($F's < 1.86$, $p's > 0.06$). Error bars represent SEM. n = 120 cells collected from 8 independent rats. (**b**) Relationship between the change in firing to the high and low value cues and the change in free choice behavior. The difference in firing to the high and low value cues in the first (blue) and last (red) 5 trials of all the blocks is plotted against the percentage of choice of the short reward during these same trials. The two measures were strongly correlated (scatter plot) reflecting the shift in both measures from early to late (blue to red in the distribution plots) within each block (Two-sided Wilcoxon ranksum test, cue, $p < 0.01$; choice, $p < 0.01$). n = 120 cells collected from 8 independent rats. Note all models implemented in the main text also produced changes in signal to the cues that differed by value (see Extended Data Fig. 5).

https://doi.org/10.1038/s41593-023-01310-x



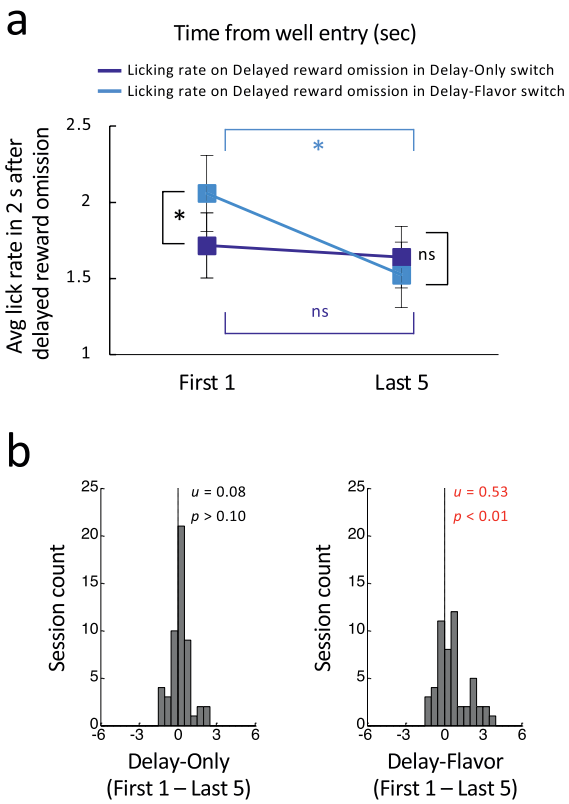
Extended Data Fig. 2 | Changes activity of reward-responsive dopamine neurons to omission of a delayed reward in delay-only and delay-flavor blocks do not depend on order of switches. Displays as in Fig. 5a of main text except that (**a**) shows data from blocks 2 and 3 data in which the delay-flavor block preceded the delay-only block, and (**b**) shows data from blocks 4 and 5 in which the delay-only block preceded the delay−flavor block. Statistics in each panel indicate results of Wilcoxon signed-rank test (p) and the average difference score (u). Comparisons of the distributions in panels a and b showed that they were not different (Two-sided Wilcoxon rank sum test) within either delay-only (p = 0.48) or delay-flavor switches (p = 0.60). n = 120 cells collected from 8 independent rats.

https://doi.org/10.1038/s41593-023-01310-x



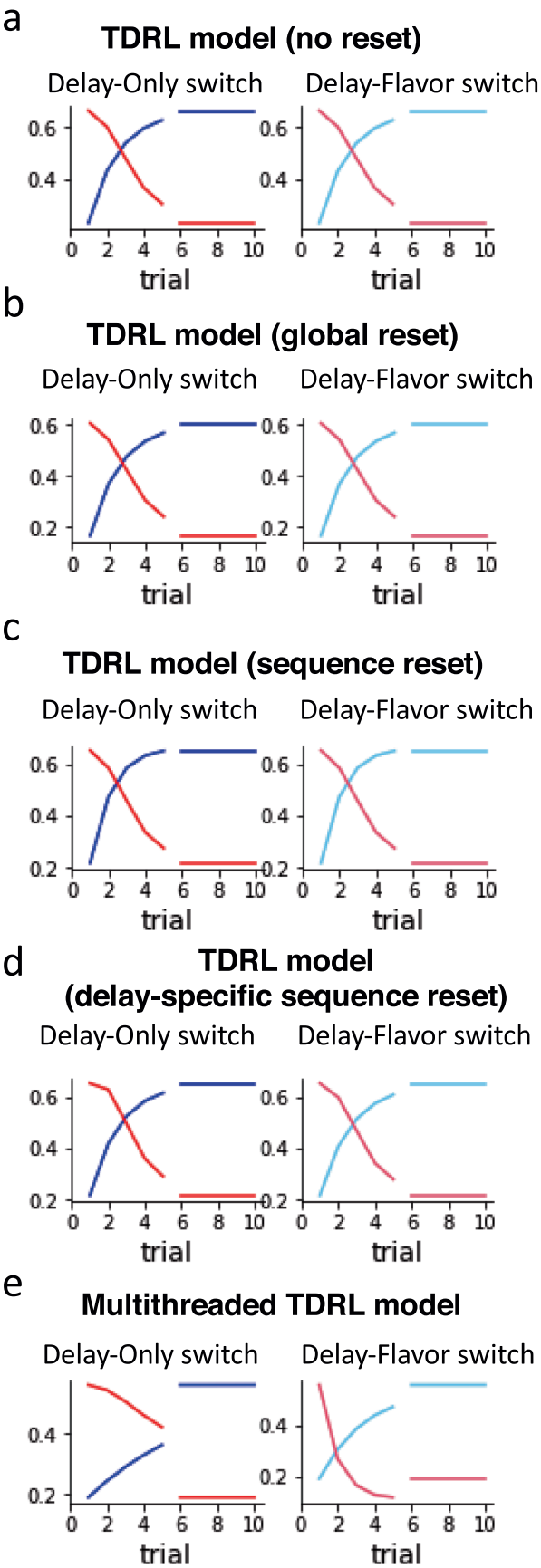
Extended Data Fig. 3 | Changes activity of reward-responsive dopamine neurons to omission of a delayed reward in delay-only and delay-flavor blocks do not depend on non-significant numerical differences in subjects' consumption of the two flavors (Fig. 1c). Displays as in Fig. 5a of main text except that (**a**) shows data involving omission of the numerically-higher reward and (**b**) show data involving omission of the numerically-lower reward. Statistics in each panel indicate results of Wilcoxon signed-rank test (p) and the average difference score (u). Comparisons of the distributions in panels a and b showed that they were not different (Two-sided Wilcoxon rank sum test) within either delay-only (p = 0.29) or delay-flavor switches (p = 0.88). n = 120 cells collected from 8 independent rats.

https://doi.org/10.1038/s41593-023-01310-x



Extended Data Fig. 4 | Licking behavior during omission of long reward. (**a**) Average lick rate in 2 sec after an omission of delayed reward on first trial and average of last 5 trials in the Delay-Only (dark-blue) and the Delay-Flavor (light-blue) switches. Two-way ANOVA (early/late x switch types) revealed a significant main effect of early/late ($F_{1,51} = 9.47$, $p < 0.01$) and a significant interaction between early/late and switch type ($F_{1,51} = 5.18$, $p < 0.05$). A step down comparison revealed that lick rates in the first trial were significantly higher than those in the last trials after a Delay-Flavor switch ($F_{1,51} = 9.90$, $p < 0.01$, light-blue), but not after a Delay-Only switch ($F_{1,51} = 0.54$, $p > 0.10$, blue). Lick rates on the first trial in the Delay-Flavor switch were significantly higher than those in the first trial after a Delay-Only switch (Two-way ANOVA, $F_{1,51} = 4.11$, $p = 0.04$), but not during the last 5 trials ($F_{1,51} = 0.85$, $p > 0.10$). Error bars represent SEM. n = 53 sessions collected from 8 independent rats. (**b**) Distributions of difference scores comparing lick rates on the first and last trials after Delay-Only (left) and Delay-Flavor (right) switches. The numbers in each panel indicate results of Two-sided Wilcoxon signed-rank test (p) and the average difference score (u). n = 53 sessions collected from 8 independent rats.

https://doi.org/10.1038/s41593-023-01310-x



Extended Data Fig. 5 | Simulated prediction error response to the cue in high and low value blocks for each TDRL model. Simulated reward prediction error responses to the cue in the (a) single thread TDRL model without reset, (b) single thread TDRL model with reset, (c) single thread TDRL model with sequential reset, (d) single thread TDRL model with delay-specific sequential reset and the (e) multithread TDRL model for each of the delay and delay-flavor block switches. All models predicted qualitatively similar changes in activity to the high value (blue) and low value (red) cues across first and last 5 trials of delay-only and delay-flavor block switches.

Geoffrey Schoenbaum
geoffrey.schoenbaum@nih.gov
Angela Langdon
alangdon@princeton.edu
Yuji Takahashi
yuji.takahashi@nih.gov

Corresponding author(s):

Last updated by author(s): Nov 24, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data was acquired using Plexon hardware (MAP Systems) and software and behavioral control was implemented using custom programs written in C++.
Data analysis	Data was analyzed using Plexon software (Offline Sorter v4) combined with custom scripts written in Matlab (v2016-2020).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The dataset and all scripts used in this study have been made publicly available as indicated by the information in the manuscript under Data and Code availability.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Study examining neural correlates (single unit firing) of errors in reward prediction. Data are quantitative.
Research sample	Subjects are Long-Evans rats of an age, sex, and number to conform to prior studies from our lab showing prediction error correlates on which this study is built.
Sampling strategy	Neural activity was acquired daily as rats completed training sessions using driveable fine-wire microelectrodes and recording on MAP systems from Plexon. Subsequently units from completed sessions were sorted and subjected to a standard waveform-based cluster analysis to separate them and classify them as putative dopamine neurons according to established procedures as described in the text using Offline Sorter v4. The numbers of rats, sessions, and ultimately neuron and dopamine neuron count were based on what is known to reveal robust error signaling in the control blocks in prior work in this task. Factors in the design were pseudorandomized by computer or experimenter where possible and appropriate; the experimenter was not blind to experimental condition.
Data collection	Data was acquired by computer using Plexon hardware and software, combined with custom programs for behavioral control written in C++.
Timing	November 2017-December 2018
Data exclusions	No data were excluded from analyses.
Non-participation	No subjects were eliminated or dropped from the analyses.
Randomization	There is only one experimental group.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Male Long-Evans rats, aged ~3 months at the start of the experiment
Wild animals	No wild animals were used in the study.
Field-collected samples	No field collected samples were used in the study.
Ethics oversight	These experiments received ethical approval from the National Institute on Drug Abuse Animal Care and Use Committee under animal study protocols 18-CNRB-108 and 20-CNRB-108.

Note that full information on the approval of the study protocol must also be provided in the manuscript.