Science
AAAS

Supplementary Materials for

Mesolimbic dopamine release conveys causal associations

Huijeong Jeong *et al*.

Corresponding author: Vijay Mohan K Namboodiri, vijaymohan.knamboodiri@ucsf.edu

Science **378**, eabq6740 (2022)
DOI: 10.1126/science.abq6740

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

Materials and methods:

Subjects:

Eight adult wild type (C57BL/6; #000664; Jackson Laboratory) mice were used for the Experiments 1-6 (six males). One male mouse was only used for Experiment 1. For Experiment 7 (sequential conditioning), seven adult DAT-Cre heterozygous mice (B6.SJL-*Slc6a3*$^{tm1.1(cre)Bkmn}$/J; #006660; Jackson Laboratory; males) and seven wild type mice (three males) were used. Data from one DAT-Cre heterozygous mouse were excluded from population analysis in Fig 6 (shown in Fig S12; see Experiment 7 in Data analysis). Additionally, six adult wild type mice were used for Fig S8L (four males). The animals were individually housed after surgery under a 12 h light/dark cycle. All experimental procedures were performed in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the UCSF Institutional Animal Care and Use Committee.

Behavioral task:

Experiment 1 (Random reward): To test the change of unpredicted reward response across experiences, a small drop of sucrose (3 µl, 15% in water) was delivered unpredictably. Animals had no prior experience of sucrose before the first session, which allowed us to measure the initial response of reward in the experimental context. The inter reward interval (IRI) was drawn from an exponential distribution with mean interval of 12 s. Animals were trained on Experiment 1 for 4-11 sessions (7 ± 2.1 sessions, mean ± std) and each session comprised of ~100 rewards.

Experiments 2-5 (Pavlovian conditioning): After Experiment 1, the same animals were trained on a discriminative Pavlovian conditioning task. A trial started with the delivery of a 2 s sound cue, either CS+ or CS- (Experiment 2). A 12 kHz continuous tone and a 3 kHz pulsed tone were used, and one of them was randomly assigned as CS+ (counterbalanced across animals). A 1 s trace interval followed the cue offset and a reward was given at the end of the trace interval only in CS+ trials. After a 3 s fixed interval from the outcome (reward for CS+ or omission for CS-; allows time for consumption), the inter trial interval (ITI) commenced. ITI was randomly drawn from a truncated exponential distribution with 30 s mean and 90 s maximum. Each session consisted of 100 trials in total (50 CS+ and 50 CS-). Animals were trained on Experiment 2 until they showed clear anticipatory licking behavior in CS+ trials. Once the anticipatory licking was high for CS+ but not CS- (*19*), at least 2 more sessions were recorded (13.3 ± 2.7 sessions in total).

The same animals went through Experiment 3, Experiment 5, and Experiment 4 chronologically. The switch from Experiment 5 to 4 was interceded by a return to Experiment 3 to train them back on the full contingency. In Experiment 3, the duration of the CS was elongated from 2 s to 8 s. Again, animals were trained until they showed stable behavior for at least 3 sessions (4.3 ± 0.8 sessions). In Experiment 5, unpredicted rewards (background rewards) were randomly delivered with a mean interval of 6 s during ITI in both CS+ and CS- trials (8.1 ± 2.5 sessions). To avoid the contamination of dopamine or behavior responses by residual response from the previous background reward, we enforced at least a 6 s delay from the last background reward to the next cue. To avoid satiation, we reduced the total trial number from 100 to 40 (20 CS+ and 20 CS-). We then retrained the animals on Experiment 3 for 2-3 sessions before transitioning to extinction. This was to restore the behavioral changes from Experiment 5. Lastly in Experiment 4, we tested extinction by omitting rewards that are predicted by CS+ (2.6 ± 0.8 sessions). After extinction and before moving to Experiment 6, animals went through 1-2 reacquisition sessions. Each session comprised of ~100 trials.

Experiment 6 (Trial-less conditioning): A brief sound cue (250 ms; same CS+ used in Experiments 2-5) was delivered unpredictably. Here, instead of defining inter-trial interval from reward to next cue, we drew inter-cue interval from truncated exponential distribution with mean interval of 33s (minimum 250 ms, maximum 99 s). Three seconds following each cue, a sucrose reward was given. With a low probability, the interval between two neighboring cues could be shorter than the cue-reward delay (3 s). These cues were defined as intermediate cues, and the paired reward as intermediate rewards.

Experiment 7 (sequential conditioning): Here, a different set of animals was used from Experiment 1-6. After one session of random reward task (Experiment 1) for lick training without optical inhibition, animals were trained on sequential conditioning. Two 1 s auditory cues (CS1 and CS2; 12 and 5 kHz tones; the identity of tones was counterbalanced across animals) were delivered in a sequence with 0.5 s interval between them. After 0.5 s trace interval from the offset of second cue, reward was given which followed an exponentially distributed ITI with 60 s mean and 180 s maximum. To optically inhibit mesolimbic dopamine throughout learning, 473 nm laser (~11mW; SLOC) was delivered to VTA from the onset of second cue until reward delivery. Animals were trained until they showed asymptotic anticipatory licking behavior (5.6 ±

2.4 and 7.1 ± 0.7 sessions for WT and DAT-Cre, respectively). Each session comprised of ~50 trials.

Based on the simulations of RPE and ANCCR, we expected to have >80% power within each individual animal to distinguish the hypotheses in most experiments. Thus, the total number of animals (n=8 for experiment 1 and n=7 for experiments 2-6) was determined to match typical sample size in the field. For experiment 7, we chose the same sample size (n=7 per group) since the distinction between RPE and ANCCR in simulations was high enough that we could test for presence of behavioral learning and dopamine cue response within each individual animal at high statistical power. The experimenters were not blinded to group identity.

Surgeries:

Surgeries were performed under aseptic conditions using procedures similar to those described previously (*19*). Animals were anesthetized with isoflurane (3-5% for induction and 1-2% for maintenance). To measure the dopamine release in the nucleus accumbens, 500 nL of dLight1.3b (AAVDJ-CAG-dLight1.3b, 2.4 x 10^{13} GC/ml, diluted 1:10) was injected in NAcc (AP 1.3, ML +/-1.4, DV -4.55) at 100 nL/min. To avoid the backflow of viruses, the injection needle was held in for additional 10 min after the completion of injection. An optic fiber (NA 0.66, 400μm, Doric Lenses) was implanted 350 μm above the virus injection site and a custom-designed head ring was implanted above the skull for the head fixation. For optical inhibition of VTA dopaminergic neurons in Experiment 7, 500 nL of Cre-dependent stGtACR2 (AAV1-hSyn1-SIO-stGtACR2-FusionRed, 1 x 10^{13} GC/ml, diluted 1:2) was injected in VTA (AP -3.2, ML ±0.99, DV -4.51 with 5° angle) bilaterally, and optic fibers (NA 0.39, 200μm, RWD) were implanted 100-200 μm above the virus injection site. After surgery, animals recovered for at least a week before water deprivation. Animal weights were logged daily during water deprivation and maintained above 82% of pre-deprivation weight (~85-88%). Daily water intake was adjusted to maintain weights at ~85-88% of pre-deprivation weight. After experiments, placement of the optical fibers and expression of viruses were confirmed using a Keyence microscope.

Photometry:

After three weeks from the virus injection, recordings were performed using either a commercial fiber photometry system (Doric Lenses) or an open-source system (PyPhotometry). In commercial system from Doric Lenses, two excitation wavelengths of LED, 405 nm (isosbestic control signal) and 470 nm (dopamine dependent signal), were sinusoidally modulated at distinct frequencies (209 Hz and 530 Hz, respectively) via the LED driver and emitted from the fluorescence mini cube. Fluorescence signals were detected using the same fluorescence mini cube at 12 kHz sampling frequency. Signals were demodulated and downsampled to 120 Hz. In PyPhotometry, two wavelengths of LED were emitted from the fluorescence mini cube (Doric Lenses) alternatively at 130 Hz using time-division sequence (*91*). Fluorescence signals were detected using a separate photodetector (Doric Lenses). Light intensity at the tip of the patch cord was maintained at 40 μW across sessions for both systems.

The 405 nm and 470 nm signals were smoothed with 200 ms window and baseline corrected using either general least square fit (*92*) or adaptive iteratively reweighted Penalized Least Squares (airPLS) algorithm (*93*). The 405 nm signal was then aligned to 470 nm signal using linear regression. Fractional fluorescence signal (ΔF/F) was defined as (470 nm signal – fitted 405 nm signal)/(fitted 405 nm signal).

Data analysis:

Experiment 1: Dopamine response to reward was defined as the normalized area under curve (AUC) of ΔF/F during reward period. Reward period was defined as -0.5 to 1 s from the first lick after reward delivery. We defined the window with reference to the first lick time, not reward delivery time, because the response latency to reward differs across trials. Also, we used a window starting from 0.5 s ahead the first lick because dopamine response started to increase even before an animal made the first lick (as they get better at sensing reward delivery in late sessions. The AUC during 1.5 s time window before reward period was subtracted from AUC during reward period to normalize baseline activity. To test dynamics of dopamine response to reward, Pearson's correlation was calculated between dopamine response and reward number, or dopamine response and inter reward interval (IRI) from the previous reward.

To test whether lick rate affects dopamine reward response, we first classified licks into consummatory and non-consummatory licks. A group of licks with less than 1 s interval was defined as a lick bout. This criterion was conservative. Every lick in the first lick bout after reward delivery was defined as a consummatory lick, and all other licks were defined as non-consummatory licks. Reward-by-reward correlation between consummatory lick rate and dopamine response was measured.

To avoid any influence of the previous reward on the dopamine response or behavior to the current reward, we excluded rewards with less than 3 s IRI from the previous reward (22.6±2.4%) for the above analyses. Rewards without lick until the next reward (5.7±0.1%) were also excluded from analyses.

To examine if changes in baseline dopamine activity caused the dynamics of reward response across sessions, baseline dopamine activity was separately analyzed from reward response. We defined baseline dopamine activity as spontaneous peaks of ΔF/F that are not reward or lick responses. A peak whose size is above 20% of the top 5% peaks of the first session and occurring >0.5 s away from any lick and >2 s from reward was considered as baseline dopamine event (i.e., related to dopamine release). The Pearson's correlation between session and averaged magnitude of baseline activity of each session was calculated for each animal, and this was compared to the correlation between session and averaged reward response.

Experiments 2-5: To analyze learning-dependent changes of behavior, the cumulative sum of anticipatory lick rate across trials was calculated. Lick rate during the baseline period (1 s window before CS+ onset) was subtracted from lick rate during 2 s from CS+ onset. Similarly, to analyze learning-dependent dynamics of dopamine response, AUC of ΔF/F during 2 s from CS+ onset was normalized by AUC during baseline period. For Experiment 3, instead of lick rate, the median time of lick onset during cue and trace periods was used to measure the timing onset of anticipatory licking. A trial in which the cumulative sum of behavioral or dopamine response was farthest from the diagonal was defined as the 'change trial', and the distance from the diagonal in the change trial was defined as the 'abruptness of change'.

In Experiment 3, two different models, one assuming a gradual change of response abiding by a Weibull function over learning (changing model) and the other assuming no change (constant model), were fitted to behavioral and dopamine responses to check whether learning caused a significant change in behavioral and dopamine responses. The equation of each model is shown as below:

Changing model:

$$\text{Response} = A\left(1 - 2^{-\left(\frac{trial}{L}\right)^2}\right) + b$$

where *A, L* and *b* were free parameters.

Constant model: $\text{Response} = 0 \times trial + b$

where b was a free parameter.

For model comparison, the Akaike information criterion (AIC) of each model was calculated as: $AIC = 2k + n\ln\left(RSS\right)$

Where k is the number of parameters, n is number of trials, and RSS is the residual sum of squares.

To test if the omission response appears when cue duration changes from 2 s to 8 s (Experiment 3), in 8 s cue trials, we calculated baseline subtracted AUC of ΔF/F for 3 to 4 s from cue onset, which was the reward period before the cue duration change. This was compared to the omission response in extinction (Experiment 4), in which we used the baseline subtracted AUC of ΔF/F for 1 s after the offset of trace interval (i.e., the original reward time).

To test if dopamine response shifts backward from reward to cue over the course of learning (Test 9), we calculated baseline subtracted AUC of ΔF/F during 1 s windows after cue onset (early) or before reward delivery (late) in first 400 trials (8 sessions). To analyze population data, we normalized early and late AUC by averaged early response during last 50 trials for each animal.

Experiment 6: We compared dopamine response to intermediate cue/reward (see Behavioral task) and previous cue/reward. Dopamine responses to cue and reward were defined as baseline-subtracted AUC of ΔF/F for 0.5 s window after cue and the first lick since reward, respectively. Baseline activity was normalized by subtracting the AUC of ΔF/F during 0.5 s window before the previous cue onset. If the interval from intermediate cue to the previous cue or to the next reward (paired to previous cue) is shorter than 0.5 s, those were excluded from analysis to prevent overlap of the analysis windows. Only sessions after the animals began to exhibit significant anticipatory licking behavior were used for analysis. For the analysis

of reward response, only rewards with the first lick before the next reward were used.

To rule out any influence of the remaining response of the previous cue on the intermediate cue response, we further examined the peak ΔF/F following each cue after subtracting the ΔF/F immediately before the cue delivery.

Experiment 7: Similar to the analysis for Test 9, baseline subtracted AUC of ΔF/F during CS1 or CS2 was calculated in first 200 trials (8 sessions) and normalized by averaged CS1 response during last 50 trials for each animal. To measure the level of dopamine inhibition, baseline subtracted AUC of ΔF/F during CS2 or reward was calculated on the first day of training and compared between DAT-Cre and WT groups. For CS2 response, AUC was calculated from -0.5 s to 2 s since CS2 onset with 0.5-s bin size and normalized by reward response in the same session for each animal. For reward response, AUC was calculated for 1 s since the first lick after reward delivery and normalized by averaged reward response from random reward session, which was performed before the first day of training. One DAT-Cre animal who showed negative reward response in the first session because of strong dopamine inhibition was excluded from population analysis in **Fig 6** (See **Fig S12**). To show the learning curve, baseline subtracted AUC of ΔF/F for 1.5 s from CS1 was measured for first three and last two sessions. Similarly, baseline subtracted number of anticipatory licks were calculated from CS1 until either reward (**Fig 6**) or CS2 (**Fig S12**).

Models:

Model 1. Prospective temporal difference reinforcement learning (TDRL): In TDRL, animals are assumed to calculate a value (V) in every time stamp t to predict reward. Time since event x was represented as an activation of different set of states and value was expressed as a weighted sum of these states.

$$V_t(x) = w_t^T x_t = \sum_{i=1}^{m} w_t(i)\, x_t(i) \tag{1}$$

where m is the total number of states of event x and $x_t(i)$ is the activation of the ith state of event x (*94*). Here, we used a linear TD(λ) algorithm.

When the prediction from the previous moment is different from the sum of current experience and discounted prediction, a reward prediction error (RPE, δ) occurs.

$$\delta_t = r_t + \gamma V_t(x_t) - V_t(x_{t-1}) \tag{2}$$

Where γ is a temporal discounting parameter.

The weight of each state is updated by RPE. Here, e represents the eligibility trace, which reflects how modifiable each state is at a given moment. An eligibility trace of each state is temporally discounted by γ, a temporal discounting parameter as above, and λ, a factor that determines the eligibility for update of any given state.

$$w(i)_{t+1} = w(i)_t + \alpha\delta_t e(i)_t \tag{3}$$

$$e(i)_{t+1} = \gamma\lambda e(i)_t + x(i)_t \tag{4}$$

TDRL models can be further divided into two types based on how the states are represented. We used two different models, complete serial compound (CSC) and microstimulus models (*94*). CSC model assumes that every moment has a separate state that is completely distinguishable from neighboring states.

On the other hand, microstimulus model assumes a temporal generalization by assuming Gaussian function of states.

$$y_{t+1} = y_t d \tag{5}$$

$$x_t(i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(y_t - \frac{i}{m} \right)^2}{2\sigma^2} \right) \tag{6}$$

Here, y refers the memory trace, which decays exponentially by a decay parameter (d), σ refers the width of Gaussian function, and m refers the number of states that each stimulus elicits. This allows for a wider and shallower shape of state as time since stimulus x increases.

CSC model was used as a representative model of TDRL. We used the following set of parameters unless noted. When the scale of results hugely deviated from the results of Model 2 (e.g., number of trials until acquisition being a lot larger for TDRL than Model 2), parameters of Model 1 were adjusted to allow a direct comparison of the two models, while parameters of Model 2 were kept as the same. In Model 1, to avoid the number of states being too large, we assumed a truncation of the state space when the interval from previous event reaches 3 times the inter-stimulus interval.

Parameters for CSC were state size = 0.2 s, $\alpha = 0.05$, $\gamma = 0.95$

Parameters for microstimulus were state size = 0.2 s, $\alpha = 0.05$, $\gamma = 0.95$, $\lambda = 0.95$, $m = 20$, $\sigma = 0.08$, $d = 0.99$.

Model 2. Retrospective causal learning: Retrospective causal learning revolves around identifying the causes of learned and innate meaningful causal targets. This learning occurs retrospectively by looking backward when experiencing a meaningful causal target to check if another stimulus consistently precedes it more than that expected by chance.

We assume that the history of every event is maintained as an eligibility trace, which is the sum of exponentially discounted trace of each prior occurrence of the event. An exponentially decaying eligibility trace allows the online calculation of the average rate of previous events without directly storing all the timestamps of previous events. The eligibility trace of event i can be expressed as below.

$$E_{\leftarrow i}(t) = \sum_{t_i \le t} e^{-\left(\frac{t - t_i}{T}\right)} \tag{7}$$

The decay of eligibility trace depends on the temporal decay parameter, T. For simplicity, we set T to be 1.2 times the average inter-reward interval of an environment. This allows the timescale of causal learning to be set by the rate of the rewards. In reality, the temporal decay parameter will itself need to be learned based on the inter-reward interval or selected from a pool of time constants (*65*), though we leave such considerations out for simplicity.

An event j is considered a meaningful causal target based on the following indicator function

$$\mathbb{I}(j \in MCT) = \mathbb{I}(DA_j + b_j > \theta) = \begin{cases} 1, & \text{if } DA_j + b_j > \theta \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where DA_j, the dopamine response to event j (see below) measures the learned contribution to the importance of the event and b_j measures the innate contribution to the meaning of a stimulus. θ is a threshold, set to 0.6. *MCT* refers to the set of all meaningful causal targets. In words, a stimulus becomes a meaningful causal target when the sum of the learned and innate contributions to its importance exceeds a threshold. For a reward, b_j is high enough to cross the threshold as it is innately meaningful (we set it to 1). For a cue, b_j will depend on its sensory properties (e.g., very loud tone is likely innately meaningful). However, for simplicity, we set it to 0 for all non-rewarding stimuli. The equations below will define the value of DA_j.

For any event j that is a meaningful causal target, the predecessor representation (PR, denoted by $M_{\leftarrow}$) of every other event in the state space is updated at the time of j as

$$M_{\leftarrow ij} \equiv M_{\leftarrow ij} + \alpha \left[E_{\leftarrow ij} - M_{\leftarrow ij} \right] \mathbb{I}(j \in MCT) \tag{9}$$

where $E_{\leftarrow ij} = E_{\leftarrow i}(t = t_j)$ measures the eligibility trace of *i* at the current time of *j* and $\equiv$ denotes an update operation. The update operation is usually shown by a backward arrow. However, we are avoiding this notation to prevent confusion with the arrow referring to a retrospective association. *Please note that we use the arrow prior to the subscripts in the symbols to denote that M← is a matrix. We do not follow this notation in the figures to allow ease of reading in the figures.*

The predecessor representation can be high merely because the rate of *i* in the environment is high. To calculate whether *i* precedes *j* beyond that expected by chance, we first calculate the baseline predecessor representation of *i* as (equivalent to baseline rate of *i* multiplied by *T*)

$$M_{\leftarrow i-} \equiv M_{\leftarrow i-} + k\alpha\left[E_{\leftarrow i-} - M_{\leftarrow i-}\right] \tag{10}$$

where – represents random moments. For simplicity, we assumed that it is updated continually every 0.2 s. We used the same learning rate as above, but with the possibility that the learning rate for updating the quantities sampled at random moments is a factor *k* times *α* (*k* is usually set to 1, see below).

The predecessor representation contingency (PRC, denoted by $C_{\leftarrow}$), the measure of the retrospective association, can now be calculated as

$$C_{\leftarrow ij} = M_{\leftarrow ij} - M_{\leftarrow i-} \tag{11}$$

An intuitive understanding of this quantity is that it is the average discounted number of *i*'s selectively preceding each occurrence of *j*.

As we showed in **Fig S3** and Supplementary Note 1, the prospective association, i.e., the successor representation contingency (SRC, denoted by $C_{\rightarrow}$), between *i* and *j* can now be calculated using Bayes' rule as

$$C_{\rightarrow ij} = C_{\leftarrow ij}\frac{M_{\leftarrow j-}}{M_{\leftarrow i-}} \tag{12}$$

An intuitive understanding of this quantity is that it is the average discounted number of *j*'s selectively following each occurrence of *i*.

The above equations are fully sufficient to learn all pairwise relationships between events (or states) in the environment. If there are N states in a state space (e.g., cues that could potentially predict a reward), there will be N^2 associations to be learned based on the above equations. The rest of the treatment below is to reduce the number of associations to be learned to N multiplied by the number of meaningful causal targets. Since the number of meaningful causal targets << N, this reduction will considerably save memory and computational cost. The primary goal of the remaining treatment is thus to define which stimuli become meaningful causal targets, thereby triggering retrospective causal learning.

We now define the net contingency (denoted as $C_{\leftrightarrow}$, note the double arrow) as the weighted sum of the prospective and retrospective association (i.e., the SRC and PRC) with a weight of *w* as

$$C_{\leftrightarrow ij} = wC_{\rightarrow ij} + (1-w)C_{\leftarrow ij} \tag{13}$$

The net contingency measures the raw association between *i* and *j*. If this quantity exceeds a threshold *θ*, there is a putative causal relationship between *i* and *j*. This is denoted by an indicator variable $\mathbb{I}(i \leftrightarrow j)$ calculated as below

$$\mathbb{I}(i \leftrightarrow j) = \mathbb{I}\left(C_{\leftrightarrow ij} > \theta\right) = \begin{cases} 1, & \text{if } C_{\leftrightarrow ij} > \theta \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

Next, we calculate the adjusted net contingency (ANCCR, denoted by $\hat{C}_{\leftrightarrow}$) between *i* and *j* by accounting for unexplained causes of *i* (**Fig S4**). If *i* is itself caused by another event *k*, then the net contingency between *i* and *j* is reduced by the

adjusted net contingency between *k* and *j* if *k* preceded *i* recently. The logic behind this adjustment is derived in the next section.

$$\hat{C}_{\leftrightarrow ij} = C_{\leftrightarrow ij} R_{ij} - \sum_{k \neq i} \left(\hat{C}_{\leftrightarrow kj} \Delta_{k \leftarrow i} \mathbb{I}\left(k \leftrightarrow i\right) \right) \tag{15}$$

Here, Δ measures recency of *k* with respect to *i* and is defined as

$$\Delta_{i \leftarrow j} = e^{-\left(\frac{t_j - t_i}{T}\right)} \tag{16}$$

where t_j is the timestamp of the current moment containing *j* and t_i is the previous timestamp of *i*. This is essentially a non-cumulative eligibility trace. R_{ij} refers to the causal weight attributed to *i* and *j* to account for the "magnitude" of *j*. For simplicity, we use this only when *j* is a reward (see below). In this case, R_{ij} is the causal weight given to the *ij* connection to account for the reward magnitude of *j*.

We now postulate that the dopamine response to an event *i* (denoted by DA_i) is the sum of ANCCRs of *i* with respect to all meaningful causal targets in the environment.

$$DA_i = \sum_j \hat{C}_{\leftrightarrow ij} \mathbb{I}\left(j \in MCT\right) \tag{17}$$

Next, we show the update rules governing the causal weight R_{ij}. We define R_{jj} as the externally signaled reward magnitude of *j* (denoted as $R_0(j)$)

$$R_{jj} = R_0(j) \tag{18}$$

This quantity is the reward magnitude for a current rewarding stimulus. For any other current stimulus, it is assumed to be 0. We update the causal weight for *ij* at moments of *j* using a standard delta rule as

$$R_{ij} \equiv R_{ij} + \alpha_R \delta_{ij} \tag{19}$$

Here, the prediction error δ is computed in a manner dependent on the sign of the dopamine response to *j* as

$$\delta_{ij} = \begin{cases} R_{jj} - R_{ij} & , \quad DA_j \geq 0 \\ \left(0 - R_{ij}\right) \dfrac{n_i^{-1} \Delta_{i \leftarrow j} \mathbb{I}\left(i \leftrightarrow j\right)}{\sum_{k \neq j} \left(n_k^{-1} \Delta_{k \leftarrow j} \mathbb{I}\left(k \leftrightarrow j\right) \right)} & , \quad DA_j < 0 \end{cases} \tag{20}$$

where n_i is the total number of times event *i* has been experienced, which can be calculated as the eligibility trace of *i* over an infinite time horizon as $n_i = E_{\leftarrow i}\left(T = \infty\right)$. The negative DA_j update rule requires some explanation. We show below that when DA_j is negative, that means that the causal weights R_{ij} were set to be too high, i.e., *j* was "overcaused" or "overexpected". In this case, the algorithm corrects the internal model by tuning down the causal weights R_{ij} for all recently experienced (accounted by Δ) putative causes of *j* (accounted by $\mathbb{I}\left(i \leftrightarrow j\right)$) based on an estimate of uncertainty (n_i^{-1} is proportional to the variance of the estimate of mean contingencies of *i*). In words, when *j* is overcaused, the causal weights of recent putative causes of *j* are reduced in proportion to the inverse of the number of times these causes have been experienced in the environment. When multiple putative causes were equally recently experienced, the reduction of causal weights is strongest for the most uncertain cause (i.e., least experienced).

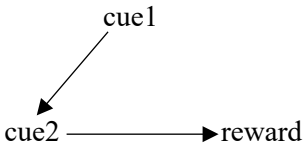
For simulations of our own experiments, we used the same set of parameters across simulations. These were *T=1.2 IRI*, α=0.02, *k=1*, *w=0.5*, and α_R=0.2. For an animal experiencing a reward for the first time, we assumed that α starts from 0.25 and exponentially decays until it reaches 0.02 with decay constant of 0.1. This was to speed up learning; a decay of learning rate is commonly assumed after a sudden change in environmental statistics (here, the introduction to an experimental

chamber), and some decay is a necessary requirement for convergence of mean estimates. For simulations of previous experiments in **Figs 2** and **S13**, we did make some adjustments to these parameters since the experimental settings were all different from each other. These are explained later.

We propose that behavior is controlled in a similar manner as for standard TDRL. Basically, value of an event is the sum of its SRC with respect to future events multiplied by their causal weights (related to reward magnitude). More generally, value might be controlled by a net contingency biased towards SRC instead of strictly determined by SRC. However, we leave this out for simplicity. This value is reduced by the cost of the action and then used to guide action selection using a softmax rule. This is explained later in the description of the simulations.

Derivation of the causality-adjustment term for ANCCR: In the previous section, we wrote equation (15) for the AN-CCR adjustment without explaining it from a fundamental principle. Here, we do so. To make things concrete, imagine the example experiment presented in **Fig S4B** (example 1). In this experiment, the animal experiences cue1 followed by cue2 followed by reward. This is also the experiment in **Fig 6C**. In the absence of any additional experience and under the assumption that an animal infers the simplest model in which both cues cause reward, the causal model inferred by the animal is:

Causal Model 2:



Here, the model states that cue1 results in cue2 and that cue2 results in reward.

We will now derive ANCCR for cue1 and cue2 based on causal model 2. To this end, we will first define the direct causal effect (DCE) of cue1 or cue2 on reward (*95*). The DCE of a cue on reward is the difference in reward due to the presence versus absence of the cue, while holding all other variables constant (including presence/absence of other cues). The DCE of cue1 on reward is zero in causal model 2. This is because cue1 does not affect reward without the presence of cue2. In other words, cue1's causal effect on reward is mediated through cue2. Hence, there is no direct causal effect of cue1 on reward while keeping the other variables (i.e., cue2) constant. On the other hand, cue2 has a positive DCE on reward since based on this causal model, presenting cue2 by itself would result in reward. DCE(c2, r) will be measured directly by the net contingency between cue2 and reward. Hence,

$$\begin{aligned} DCE(c1,r) &= 0 \\ DCE(c2,r) &= C_{\leftrightarrow c2r} R_{c2r} \end{aligned} \tag{21}$$

We postulate that the ANCCR of cue1 with respect to reward will be its net contingency with reward. This is because cue1 is itself presented without an antecedent cause but is strongly associated with a future reward. Thus, even though cue1 has zero DCE on reward, it still signals a positive ANCCR due to its causing of cue2, which is causative of reward. Hence, when cue2 is presented following cue1, a considerable portion of its DCE on reward has already been signaled by the previous presentation of cue1. To account for this, we postulate that the net sum of ANCCR of cue1 and cue2 on a given "trial" should equate the net sum of their DCEs on reward. This implies that on a trial in which cue1 is followed by cue2, the ANCCRs will respectively be:

$$\begin{aligned} \hat{C}_{\leftrightarrow c1r} &= C_{\leftrightarrow c1r} R_{c1r} \\ \hat{C}_{\leftrightarrow c2r} &= \left(C_{\leftrightarrow c2r} - C_{\leftrightarrow c1r} \right) R_{c2r} \end{aligned} \tag{22}$$

$R_{c2r} = R_{c1r} = R_0(r)$ here since the animal has had no experience other than cue1→cue2→reward in this example.

On the other hand, a trial in which cue2 is presented by itself (i.e., its DCE has not been previously signaled by cue1) would result in its ANCCR being

$$\hat{C}_{\leftrightarrow c2r} = C_{\leftrightarrow c2r} R_{c2r} \tag{23}$$

Overall, this means that the ANCCR of cue2 with respect to reward will be adjusted if cue2 is preceded by cue1, but not if cue2 is presented by itself. A mathematical formulation for this is to correct for cue1 based on an indicator variable that measures whether cue1 preceded cue2 on this "trial". We write this mathematically as:

$$\hat{C}_{\leftrightarrow c2r} = \left(C_{\leftrightarrow c2r} - \mathbb{I}_{c1\leftarrow c2} C_{\leftrightarrow c1r}\right) R_{c2r} \tag{24}$$

where the term $\mathbb{I}_{c1\leftarrow c2}$ measures whether cue1 preceded the current occurrence of cue2 on this "trial". We posit however that animals do not experience trials. So, to represent whether cue1 preceded cue2 recently, we replace $\mathbb{I}_{c1\leftarrow c2}$ with $\Delta_{c1\leftarrow c2}$.

Therefore, we get

$$\hat{C}_{\leftrightarrow c2r} = \left(C_{\leftrightarrow c2r} - \Delta_{c1\leftarrow c2} C_{\leftrightarrow c1r}\right) R_{c2r} \tag{25}$$

This is the equation shown in **Fig S4B**.

Since in this case, cue1 was identified as a cause of cue2 in the causal model, and $\hat{C}_{\leftrightarrow c1r} = C_{\leftrightarrow c1r} R_{c1r}$, the above equation can be rewritten in the same form of equation (15) as

$$\hat{C}_{\leftrightarrow c2r} = C_{\leftrightarrow c2r} R_{c2r} - \hat{C}_{\leftrightarrow c1r} \Delta_{c1\leftarrow c2} \mathbb{I}\left(c1 \leftrightarrow c2\right) \tag{26}$$

A similar logic applies for the reward response. Here, $R_{rr}=R_0(r)$, the observed reward magnitude. Both cue1 and cue2 are putative causes of reward since they both have a strong association with reward. However, cue2 is an "explained" cause of reward as cue1 itself causes cue2 (in the causal model). Therefore, when performing the subtractive adjustment, cue2's association with reward will itself need to be adjusted based on cue1 causing cue2. Overall, this can be written as

$$\hat{C}_{\leftrightarrow rr} = C_{\leftrightarrow rr} R_{rr} - \hat{C}_{\leftrightarrow c1r} \Delta_{c1\leftarrow r} \mathbb{I}\left(c1 \leftrightarrow r\right) - \hat{C}_{\leftrightarrow c2r} \Delta_{c2\leftarrow r} \mathbb{I}\left(c2 \leftrightarrow r\right) \tag{27}$$

This is exactly equation (15).

Overcausation of reward: Here, we explain why the sign of DA_{reward} reflects whether or not reward is "overcaused" in the estimated causal model of an animal. This was used in Equation (20).

In the real world, a sequence of rewards (magnitude R) can have many causes. Say that these causes are denoted by the letters c1, …, cn. Further, the reward may itself occur in a periodic fashion. Here, the reward can be loosely thought of as causing itself. In addition to such "caused" rewards, there may also be unpredictable rewards that have no apparent cause in the environment. Under the unrealistic assumption that the animal knows the exact direct causal effects of all causes and reward, the total rate of observed rewards should be equal to the sum of rate of causes multiplied by their direct causal effects on reward and the baseline rate of unpredicted rewards. This can be mathematically expressed as:

$$M_{r\_}R = \sum_{c} DCE\left(c,r\right) M_{c\_}R_{cr} + DCE\left(r,r\right) M_{r\_}R + M_{r,baseline} \tag{28}$$

where $M_{x,\text{-}}$ reflects the expected number of events x over some time horizon (proportional to true rate of x).

We know that the baseline rate of unpredicted rewards should be greater than or equal to zero, i.e., there cannot be fewer rewards in the environment than rewards accounted by their causes. This means that the following constraint must be true.

$$\left(1 - DCE\left(r,r\right)\right) M_{r\_}R - \sum_{c} DCE\left(c,r\right) M_{c\_}R_{cr} \geq 0 \tag{29}$$

When rewards are temporally predictable from each other (e.g., perfectly periodic), *1-DCE(r,r)* will be lower than 1. The estimated reward↔reward net contingency will also be less than 1 in this case, but will be larger than *1-DCE(r,r)* (for a perfectly periodic reward sequence, $C_{\leftrightarrow rr}$ will be 0.5 for infinite T, while *1-DCE(r,r)* will be zero). Thus, for an animal estimating these quantities, reward will be "overcaused" when

$$C_{\leftrightarrow rr}M_{\leftarrow r-}R-\sum_{c}DCE\left(c,r\right)M_{c-}R_{cr}<0 \tag{30}$$

Now suppose that one can learn the ANCCRs between the causes and reward in such a manner that the following relationship is satisfied at reward times

$$\sum_{c}DCE\left(c,r\right)M_{c-}R_{cr}\le\sum_{i}\hat{C}_{\leftrightarrow ir}M_{\leftarrow r-}\Delta_{i\leftarrow r}\mathbb{I}\left(i\leftrightarrow r\right) \tag{31}$$

Here, the sum on the R.H.S. is carried over all events *i* and not just the causes. However, because the sum checks whether each *i* is a putative cause of reward ($\mathbb{I}\left(i\leftrightarrow r\right)$), this is equivalent to summing over causes under the assumption that the putative causes reflect true causes. For the above equation to be satisfied, the animal will need to choose an appropriate *w*. When temporal discounting is not a factor and there are only unexplained causes of reward, setting *w=0* satisfies the equality because $\hat{C}_{\leftrightarrow ir}M_{\leftarrow r-}=C_{\leftrightarrow ir}M_{\leftarrow r-}=C_{\leftarrow ir}M_{\leftarrow r-}=C_{\rightarrow ir}M_{\leftarrow i-}=DCE\left(c,r\right)M_{c-}R_{cr}$. Thus, the net contingency is biased towards PRC in this case.

Substituting the constraint from Equation (31) into Equation (30), a reward is overcaused when

$$C_{\leftrightarrow rr}M_{\leftarrow r-}R-\sum_{i}\hat{C}_{\leftrightarrow ir}M_{\leftarrow r-}\Delta_{i\leftarrow r}\mathbb{I}\left(i\leftrightarrow r\right)<0 \tag{32}$$

Removing $M_{\leftarrow r-}$, we get

$$C_{\leftrightarrow rr}R-\sum_{i}\hat{C}_{\leftrightarrow ir}\Delta_{i\leftarrow r}\mathbb{I}\left(i\leftrightarrow r\right)<0 \tag{33}$$

The L.H.S. here is $\hat{C}_{\leftrightarrow rr}$, which is less than or equal to the dopamine response at reward time (equal if reward itself does not cause other rewarding meaningful causal targets). Thus, a reward is overcaused when

$$DA_{r}<0 \tag{34}$$

In Equation (20), we check this relationship for every occurrence of reward and reduce the causal weights of recent causes when it is met.

Simulations:

Experiments: We simulated all the behavioral experiments (Experiments 1-7) using Models 1 and 2. Experiment 1 was simulated with 50,000, 20,000 and 2,000 trials for CSC (Model 1), microstimulus (Model 1) and Model 2, respectively. This was determined based on the learning speed of each model. For microstimulus, α = 0.02, γ = 0.98 were used. As we did in the experimental data analysis, any trials with <3s IRI from the previous reward were excluded from the analysis. For the examination of the relationship between predicted dopamine response and trial during initial learning, predicted dopamine response across trials were fitted to a Weibull function to find an asymptote and only trials until 95% of asymptote were used for the analysis. This was repeated 100 times. Increasing *T* from 1.2 to 10 times of IRI with keeping all other parameters as the same was enough to fit the result of animal 2 (**Fig S8J**).

For the simulation of Experiments 2-6 using Model 2, we trained a model on Experiment 1 beforehand to mimic the behavioral experimental procedure. For the simulation of Experiment 3, a hypothetical experiment with short ITI (mean of 6 s) was also tested to reproduce reported scaling effect of dopamine signal with cue-reward delay. For the simulation of Experiment 5, we also tested Model 1 without ITI states, which is generally assumed in TDRL simulations. Each iteration was comprised of 2,000 trials of CS+ and CS- each. For Experiments 3-6, where the condition changed in the middle of a session, additional 2,000 trials of CS+ and CS- were tested with the changed condition. This was repeated 100 times for each simulation. For the simulation of Experiment 6 using Model 2, we trained a model on Experiment 1 beforehand to mimic the behavioral experimental procedure. Each simulation comprised of 1,000 cues and was repeated 100 times. Experiment 7 was simulated with 800 and 400 trials for Model 1 and Model 2, respectively, according to the learning speed of each model. The level of dopamine inhibition was set to constant -0.6 (**Fig 6**) or linearly decreasing value from 0 to -0.6 (**Fig S12**), which

was measured experimentally, from cue onset until reward. Model 1 was tested with variable range of lambda (0, 0.5, and 0.95; 0.95 in **Fig 6**). To test Models 1 and 2 in various conditions beyond the behavioral experiments we performed, we simulated several thought experiments and previously reported experiments.

Thought experiments:

In **Fig S6**, to test if Model 2 can learn causal structure in complex situations, we simulated three different thought experiments. First, we tested whether it could learn causal structure across different cue-reward intervals. A cue was randomly given with the average interval of 12 s and a reward was delivered after a certain amount of delay from each cue. We tested seven different durations of cue-reward delay, which spanned from 0.05 to 5 times of inter-cue interval. As the duration of cue-reward delay gets longer, the probability of having different events between paired cue and reward increases. As a control, we simulated reward and cue delivered at the same rate as above, but without an association. Next, we tested whether causal associations can be learned if multiple associations with variable delays happen simultaneously. We increased T of Model 2 to 5 times of IRI for **Fig S6A-B**, so that the association with longer delay can be learned better. Next, three cues with different durations of cue-reward delay (1, 10 and 100 s) were used with an averaged inter cue interval of 50 s. Finally, we tested a more complex situation with large number of associations. Three different associations existed, cue1→reward1, cue2→reward2, cue3→cue4→reward, with four extra cues that are not associated with anything. Here, the interval between any associated events was 2 s. For the simulations in **Fig S6**, we took an exact mean of eligibility traces of event *i* over event *j* to calculate $M_{i \leftarrow j}$, instead of updating it using a constant alpha as we were more interested in the accuracy of the model rather than the dynamics of a leaky integration. Each simulation was repeated 100 times and each iteration was continued until the number of any cue in each simulation reached 1,000.

Previously reported experiments:

In **Fig 2**, we used Model 2 to simulate a range of classic results that are predicted by temporal difference RPE models. All simulations in **Fig 2** were repeated for 20 iterations. In **Fig2A-C**, we simulated dopamine response during Pavlovian conditioning task with different reward probabilities and magnitude. First in **Fig2A**, reward was predicted 1 s after cue, and ITI was drawn from a truncated exponential distribution with 30 s mean. A total of 500 trials were tested, and early and late trials were defined as the first and the last five trials. Additionally, 100 trials were tested with 50% reward omission. We used k=0.01 to have a stable estimation of baseline event rate for the simulations in **Fig 2A, B, E, F**. In Model 2, reward omission was itself regarded as a learned state (since animals have to infer omission and stop reward seeking once they realize that omission is possible in the task) and T was set to IRI. We assumed that the animal has probabilistic access to the moment of reward omission which is scaled by reward probability. A more general mechanism should account for the temporal uncertainty of animals to infer when a reward was omitted. Since such a process must itself be learned, an implication is that animals treat reward omissions as a "state" only after the learning of such inference. This is consistent with the previous observation that negative omission responses of dopamine neurons appear only well after animals learn cue-reward associations (*54*). Omission response was calculated as the net contingency between omission state and reward. In **Fig 2B**, to test dopamine response during Pavlovian conditioning with probabilistic reward, rewards were delivered in a probabilistic manner from the start of the simulation (either 30%, 60%, or 90%) 0.5 s after cue delivery (with mean ITI of 100 seconds capped at 300 seconds). We used w=0.6 here to emphasize the difference across reward probabilities by slightly biasing model towards prospective association in **Fig 2A, B**. In **Fig 2C,** we simulated dopamine response to reward with trial-by-trial changes in reward magnitude (*79*). In 90% of trials, cue was followed by reward with reward of magnitude 5, while in the remaining 10% of trials, reward magnitude was changed to either 1 or 10.

All further simulations in **Fig 2** approximated the probability of behavioral responses to cue (or without cue in **Fig 2H**) by applying a softmax function.

$$p\left(action \mid cue_i\right) = \frac{e^{\frac{V_i}{T}}}{\sum_j e^{\frac{V_j}{T}}} \tag{35}$$

where V_i is value of cue *i*, T is temperature. We defined value of cue (V_i) as a product of the prospective association between cue and reward ($C_{\rightarrow ir} R_{ir}$) and subtracted the cost of action. The cost of action was adjusted across simulations to produce

different behaviors for cues with different values. We set the value of a null action (no lick, V_0) as 0. Here, we directly used value of cue to estimate action probabilities for simplicity, but in reality, action selection mechanism will also depend on a separate learning of temporal delay.

In **Fig 2D,** we trained a cue as a conditioned inhibitor (CI) by simulating unpredictable rewards with a mean interval of 5 seconds and then instituting a 10 s pause in reward delivery following cue presentation. Another cue (CS) was separately trained to predict a reward with 3 s delay. Both CS and CI were trained for 1000 times of presentation with 30 s averaged ITI. After initial training of the CS and CI, the probability of lick was measured during simultaneous presentation of CS and CI three times, and this was compared to the lick probability measured from last three trials of initial CS learning. In these calculations, the meaningful causal target threshold was increased to 0.8 to prevent aberrant learning of a causal association from reward to the CI. The cost of action was set to zero. In **Fig 2E**, to simulate blocking, the probability of lick to a conditioned stimulus (CS2) was estimated when learning was attempted during the presence of previously learned conditioned stimulus (CS1) and when no other stimuli were present. After 4000 trials of CS1-reward pairing, CS1 and CS2 were simultaneously presented, followed by reward. The delay between cues and reward was maintained as 0.5 s. After 300 co-presentations of CS1 and CS2, CS2 was delivered alone three times without either CS1 or reward to measure the probability of lick to CS2. As a control, we independently measured lick probability after 300 presentations of CS2 and reward alone without prior CS1-reward training. The mean of ITI was 100 s. The cost of action was set to 0.3 in **Fig 2E,F**. In **Fig 2F**, we simulated the behavior to overexpectation of rewards due to co-presentation of two learned CSs. We first independently trained associations between cues (CS1 and CS2) and a reward. Both cues predicted rewards with a delay of 0.5s and mean ITI of 30. After training of 500 trials per each cue, CS1 and CS2 were presented together, but simultaneous presentation resulted in delivery of a single reward. This compound training was repeated for 500 trials. Probability of lick to CS1 was measured three times without following reward delivery before and after compound training.

Finally, we ran a series of experiments to see how manipulation of dopamine in the ANCCR model would impact behavior. In **Fig 2G**, we simulated extinction of a behavioral response to a learned CS by inhibiting dopamine during reward delivery (*14*). A reward was delivered after 1 s from cue, and ITI with an average of 30 s was followed. After 400 trials, we substituted the value of dopamine response at reward time with -0.5 to simulate dopamine inhibition. The cost of action was set to 0.5. Next in **Fig 2H**, we inhibited dopamine at reward time to test if that can suppress the action on the subsequent trial as reported (*80*). In this simulation, the agent could perform either Action 1 or Action 2 which predicts reward with 90% or 10% probability, respectively. After 500 trials, we inhibited dopamine response to -1 at reward delivery following choice of Action 1 in every 500 trials. This was repeated 16 times. The probability of choosing Action 1 on the inhibited trial (n) (assuming could have been on any trial and not just trials with Action 1) was compared to that on the subsequent trial ($n+1$). For this simulation, we increased the value of the reward learning rate (α_R) to 0.85 to allow rapid change of behavior in trial-by-trial manner. This is in line with experimental observations (*80*). Lastly, in **Fig 2I** we demonstrated that DA stimulation within the ANCCR model can prevent the effects of blocking, which is known as 'unblocking' (*14*). The behavior simulated here was the same as the standard blocking experiment shown in **Fig 2E**, but during simultaneous presentation of CS1 and CS2, dopamine was stimulated to 1 at reward time.

In **Fig S5**, known scaling laws of behavioral learning (*82*) were tested. First, we tested the dependency of learning on the ratio of cue-reward delay and ITI. Eight different durations of cue-reward delay between 3 s and 10 s were tested. ITI was drawn from a truncated exponential distribution, but the average of ITI was either fixed at 30 s (fixed ITI) or scaled as 10 times of cue-reward delay (scaled ITI). A total of 16 conditions were tested (8 cue-reward delay lengths x 2 ITI types). We assumed that acquisition happens when the moving average of predicted cue response with 30 trials window crosses an arbitrary threshold. For a direct comparison of the two models, threshold was chosen separately for each model so that trials to acquisition from two models were comparable (0.08 and 0.6 for Model 1 and Model 2, respectively). We used state size = 1s and $\alpha = 0.025$ for Model 1 simulation. Second, we tested the dependency of learning on reward probability. Ten different reward probabilities from 10% to 100% were tested. Reward was followed with a given probability 1 s after cue, and ITI with an average of 10 s was used. Threshold was chosen as 0.08 and 0.6 for Model 1 and Model 2, respectively. State size = 1 and $\alpha = 0.001$ were used for Model 1 simulation. Each simulation was repeated 100 times.

In **Fig S13**, we simulated two recent findings using Model 2. To simulate ramping activity that arises when a series of cues predicts reward (*58*), we used 8 different cues with 1 s duration and a reward at 9 s from the onset of first cue. ITI with an average of 9 s was used. After training a model with 2000 trials of the standard condition, we tested it for either teleport or speed change condition with additional 2000 trials. For the teleport simulation, we randomly interleaved the following three

types of conditions with the standard condition. When the 2nd, 4th, or 6th cue was delivered, we skipped the next cue and teleported the model to the cue after next cue. Each test condition comprised 1% of test trials. For speed change, we randomly changed the speed of the task to 0.5X or 2X of the standard condition in 10% of test trials each. Because of the exponential decay of eligibility trace, if the delay between two associated events is shorter, the net contingency between them would be larger than the same association with longer delay. As animals repeatedly experienced trials with the three different speeds, we assumed that animals can notice the speed change quickly and adjust their net contingency and ANCCR by multiplying the speed of each trial relative to the speed of standard trials (to estimate that the reward will be available sooner or later depending on the speed). This will result in the adjustment of dopamine response. As the timescale of integration is likely be shorter in an environment with more variability compared to a stable environment, we assumed that T exponentially decays from 1.2*IRI to 0.5*IRI once trials with either teleportation or speed change are added. Threshold was set to 0.5.

Supplementary Notes

Note 1: Bayes' rule in a continuous event-based timeline:

In a continuous event-based timeline, the successor representation between cues and rewards is the mean (across cue occurrences) of the discounted sum of future reward occurrences. This can be defined mathematically as (**Fig S3**)

$$M_{\rightarrow cr} = \frac{\sum_{c=1}^{n_c} \sum_{r=1}^{n_{c\rightarrow r}} e^{-\left(\frac{\Delta t_{cr}}{T}\right)}}{n_c}$$

Where the first sum is to average over the cue occurrences, and the second sum is to average over all the reward occurrences that follow each cue occurrence. Δt_{cr} refers to the difference in time between a cue and an occurrence of a future reward, and T is the temporal discounting timescale. Please note that $M_{\rightarrow cr}$ is the cr element of the $M_{\rightarrow}$ matrix, i.e., the successor representation matrix between all states in the world.

If one assumes that every stimulus evokes an eligibility trace with a time constant of T, one can similarly define a predecessor representation between cue and reward as

$$M_{\leftarrow cr} = \frac{\sum_{r=1}^{n_r} \sum_{c=1}^{n_{c\leftarrow r}} e^{-\left(\frac{\Delta t_{cr}}{T}\right)}}{n_r}$$

Where the first sum is to average over the reward occurrences, and the second sum is to average over all the cue occurrences that precede each reward occurrence. In both cases, the exponentially discounted delay between every cue-reward pair wherein cue precedes reward is summed over in the numerator. Thus, the numerators of both the above equations are identical if one assumes that the environment is stationary. When averaging over stochastic instantiations of a stationary environment, the denominators simply refer to the expected number of cues and rewards within a given time period. These are directly proportional to the cue and reward rate and their ratio can be estimated by the ratio of the average baseline eligibility trace of cues and the average eligibility trace of rewards. This is the equation shown in **Fig S3**.

At a big picture level, this mathematical exercise allows a simple mechanism to learn the predecessor representation, and thereby calculate the successor representation. Due to the nature of exponential reductions of eligibility traces, there is no need to store the times of all previous events in memory directly. Instead, storing an eligibility trace that decays exponentially and gets added by 1 every time the cue occurs, is sufficient to get a "discounted" number of cues preceding a given reward. After conversion to the successor representation, the time constant of the eligibility trace decay determines the prospective discounting time constant.

Note 2: How about a semi-Markov state space that does not break delays into internal states?

In the introduction, we stated that TDRL models typically assume that the delay periods are themselves broken up into a sequence of internal states that do not directly reflect external sensory signals. One formal conception where this is not true is the semi-Markov state space that has been previously used to better explain the timing phenomena underlying dopaminergic signaling (*36*). In this state space, delay periods do not themselves get broken into states. Instead, it is assumed that the delay distributions are separately learned.

The semi-Markov model makes two strong predictions that are inconsistent with our data in relation to Tests 1 and 2. The model predicts that during a random rewards experiment (Experiment 1) in which rewards are delivered in a Poisson train, the RPE early in experience will be proportional to the reward magnitude, but later in learning will be proportional to the reward magnitude multiplied by 1-previous IRI/mean IRI (equation 4.4 in Section 4.3.1 of (*36*)). This makes two strong predictions: 1) dopamine responses at reward time will reduce over repeated experience of sucrose (related to Test 1), and 2) dopamine responses to sucrose will be lower following a longer IRI (related to Test 2). Both these predictions are the same as those listed under the RPE models in the main text. However, unlike regular RPE models which often do not explicitly consider the states present during an IRI/ITI epoch, the semi-Markov model provides a formal account of such periods. Indeed, due to these predictions being explicitly listed in the original paper and being violated in our data, we consider this as a ruling out of the semi-Markov model. Hence, we did not consider this model further in relation to the other tests.

Nevertheless, it is worth pointing out one apparent similarity between the semi-Markov model and our model. The

assumption in the semi-Markov model that a delay distribution between events is learned separately from the value learning system is similar to that in our model. However, there is a critical qualitative difference between the two models. While the semi-Markov model assumes that some system in the brain has learned the delay distributions *first* and conveys it to the value learning system of which mesolimbic dopamine is a part, we propose that delay distributions are learned *after* the associative strength is first learned between a cue and reward (**Fig S2**).

It is also important to note that even though we introduced the Test 1 RPE prediction by postulating that the context acquires value, contextual learning is not necessary for this prediction. So long as the animals represent the inter-reward interval as a state (or many states, as simulated in **Fig 3**), RPE will reduce with repeated exposures to sucrose. Further, in Test 2, one may initially think that either a shortening or lengthening of the IRI should induce an equivalently large prediction error. However, the RPE measures value prediction errors; shortening is an unexpected increase in value and hence, produces larger RPEs.

Note 3: Can stress explain the increase in unpredicted reward response with repeated experience?

An alternative hypothesis for the change in sucrose response with repeated experience might be that the stress from head-fixation reduces the motivation for sucrose early in training. Stress is very unlikely to explain the increase in reward responses in the random rewards experiment (Experiment 1) for a few reasons:

1) Licking is a direct measurement of whether the animals are motivated to consume sucrose. We observed little to no change in the consummatory or non-consummatory lick rate of animals across sessions, suggesting that any stress due to head-fixation does not alter the incentive value of the sucrose reward. If anything, we observed a decrease in consummatory licking with repeated experience with sucrose, a result that is not consistent with a possible explanation related to a gradual reduction in stress. Indeed, the presence of stress is often measured using sucrose consumption assays.

2) 15% sucrose is highly rewarding and it is well known that rewards such as palatable foods reduce stress (*84*).

3) RPE measurements are routinely made under head-fixation (*4*, *43*, *54*, *55*).

4) Restraint stress is known to cause an increase in baseline dopamine (*85*). However, the baseline change observed by us did not correlate with the increase in reward response (**Fig S8**).

5) A previous publication found that when animals are not pre-exposed to rewards in the experimental context, there is an initial increase in dopaminergic reward responses during Pavlovian conditioning despite extensive prior habituation to head-fixation (*54*). This is fully consistent with our algorithm and experimental results (**Fig S12**).

6) Acute stress does not alter the reward response of dopamine release in nucleus accumbens and enhances release of dopamine in the ventrolateral striatum (*86*).

Together, these results make it very unlikely that the increase in sucrose responses of dopamine release in the nucleus accumbens is due to a gradual reduction in stress from head-fixation. Further, it is important to point out that the random rewards experiment in which unpredicted sucrose rewards were delivered with high temporal precision is very challenging to perform in a freely moving preparation since the animals will not consume the rewards at the intended times. Though intra-oral cannulation may allow precise control of reward times, it is by itself a stressful procedure and does not allow ground truth validation of motivation to consume sucrose, as is available in our head-fixed preparation. Thus, our experimental preparation uniquely allowed us to perform tests 1 and 2.

Note 4: Alternative prediction error models:

One possible explanation for the observation that dopamine response to sucrose increases when the previous inter-reward interval increases (**Fig 2E**) is that dopamine signals a "reward rate prediction error". In other words, if an animal continually calculates the base rate of rewards in the environment using an online update delta rule (*78*), the delta (i.e., difference between current reward rate and the average) can be thought of as a prediction error. This quantity will be larger in response to sucrose after a long IRI since the average rate is lower. Though this may sound semantically different from our proposal, the calculation of the ANCCR is almost identical to the above calculation for the random rewards experiment. Indeed, the online updating of PRC is done using a prediction error like update (Methods; see next paragraph). That said, the reward rate prediction error hypothesis does not match the rest of the data (e.g., extinction would be expected to produce zero dopamine cue responses very quickly).

A more general observation is that the online calculation of the average (or any quantile/expectile) of samples received one at a time can be performed using a simple prediction error type update. This rule is: *new mean = old mean + step size* *

(new sample – old mean), where if step size = 1/number of samples, the update calculates the mean exactly (usually, step size = alpha to allow for non-stationarity of the distribution). This is the same rule that we use to update predecessor representations. Hence, the idea of prediction errors is much more general than the specific form of prediction error hypothesized by TDRL for associative learning. In other words, our argument here is against dopaminergic encoding of the specific form of prediction error used in TDRL (*prediction error = reward(t) + γValue(state at time t)-Value(state at time t-1)*), which is the dominant theory of dopamine's role in learning.

Further, in our model, the learned quantities measure retrospective signals and not prospective signals, even though the prospective prediction can be derived from the learned retrospective signals. On the other hand, TDRL RPE directly learns a prospective value signal. In this sense, our model is more related to model-based learning wherein an internal causal model of the world is learned by the learning agent. A key difference with current model-based approaches is that our approach does not require the construction of a state space during delay periods that obey the Markov property. Instead, our approach works when the Markov assumption holds in a causal graph (e.g., the cue-reward association is causally Markov in **Fig S6** even though it is hard to specify in a Markov state space that breaks up delay periods; see **Fig S15B** for this exact Markov state space). Future work is required to extend our approach to the learning of a more complex causal graph involving sequences of states (e.g., (*87*, *88*)).

Note 5: If behavior requires threshold crossing of net contingency, why not assume that it requires threshold crossing of TDRL value?

A potential approach within TDRL RPE to explain the observation that behavioral learning is slower and much more abrupt than the evolution of the dopaminergic cue response is to assume that the value signal learned by TDRL needs to cross a threshold (e.g., through a highly non-linear sigmoid transform) to produce the behavioral output. The softmax action selection function in TDRL in which value is compared against cost of performing the action is a potential means to obtain such thresholding. However, there are two problems with this approach.

First, the dopaminergic cue response was near its maximum (>93%) before the emergence of behavioral responses (**Fig S10C**). If this results purely from a non-linear transformation of the decision variable (i.e., value), the threshold for converting value to behavior needs to be very high. Such a high value would imply that even slight reductions in the learned value signal in the brain would result in no behavioral output, thereby implying that animals would not be able to learn cue-reward associations under partial reinforcement. Indeed, we had to assume a low threshold for value crossing to simulate the dependence of initial learning on partial reinforcement in **Fig S5B**. Here, for simplicity in Fig S5, we just assumed that behavior was generated when dopamine cue response in either model crossed a threshold. Accordingly, we had to set the threshold to be low for TDRL to cross it and result in behavioral learning for most partial reinforcement schedules. Thus, an assumption of threshold crossing of a TDRL value signal is not easily reconciled with the extensive evidence that animals can learn cue-reward associations under partial reinforcement, and our observation that dopamine cue response is already near its maximum by the time behavior appears.

Second, the entire point of directly learning and storing a value signal using model-free RL is to directly store the decision variable. On the other hand, the retrospective causal learning approach is a form of model-based learning, wherein an internal model of the world needs to be constructed first. It is common to assume some statistical threshold to test whether a given hypothesis about the world is deemed real (e.g., the 5% alpha hypothesis for statistical testing). Hence, such an assumption is also reasonable for a neural computation. Thus, the notion of a statistical boundary is more reasonable for the causal learning approach than for model-free TDRL. Further, the valuation process underlying behavioral output might not be driven by an integrated economic value signal as assumed by model-free TDRL, and may instead depend on heuristics operationalized on a world model (*89*).

Note 6: Learning cue-reward associations without loss of generality:

Common experiments testing cue-reward learning enforce a constraint on experimental design: the next cue following a given cue can only occur following the outcome of the current cue (reward or omission). This is an implicit assumption of a "trial-based" view of learning, i.e., that each cue presentation defines a trial, with the outcome of that cue (i.e., trial) fully evident before the end of the trial. This implicit assumption is what allows the construction of a state space that breaks delays into unobserved internal states (*21*). A consequence of this is that the state representation following a stimulus should reset every time the stimulus is presented again (i.e., next trial onset). In other words, after every "trial" (i.e., cue presentation), the internally triggered state space will reset to measure the passage of time from this new stimulus. Such learning fails for even simple scenarios in which a cue predicts a reward at a fixed delay with 100% probability, when the common experimentally enforced constraint is lifted (**Fig S6**). To illustrate why this approach fails, imagine an experiment in which a cue predicts reward at 10% probability in the usual trial design (i.e., next cue occurs only after the current trial's outcome

(reward or omission)). In this environment, an animal will experience many cues, but only 10% of them are followed by a reward *before the next presentation of the cue*. The last phrase of the previous sentence is the critical assumption that allows the initial learning of the fact that 90% of the trials have reward omission. Essentially, this means that to learn the reward probability as 10%, the "trial-based" models evaluate the outcome of a cue *before the next cue occurrence*, which causes the state space to be reset on every cue. Under this simplifying trial-based assumption, if the next cue occurs during a cue-reward delay, learning is affected. This is the assumption that we tested and ruled out in **Fig 4**. A possible extension is to define the state space in a manner that includes the information of the previous cue presentation. However, this too is not a general solution since it depends on the assumption that either the current or the previous cue must determine reward. Further, prior postulates of partially observable Markov decision processes based on belief states are still "trial-based" and hence, reset at cue presentation (*58*). Besides, learning in such partially observable Markov decision processes is much more complex than the simple contingency/causality-based rule proposed here. We do present the exact non-resetting state space that recovers the Markov dynamics of this environment in **Fig S15**. However, this state space requires the animal to *a priori* know the exact structure of the task, which is, after all, the goal of the learning. Hence, the retrospective causal learning model is simpler and parsimonious.

Note 7: Aren't there near-infinite cues to consider even in the computation of the retrospective association?

In the introduction, we stated that "Learning what cues precede ice-cream (retrospective) is much easier than learning the future outcome of a near-infinite number of cues, only a few of which may result in ice-cream (prospective)." On a closer inspection, this assertion may be questionable since it may not be clear why a near-infinite number of cues need not be considered for evaluating whether a cue precedes the reward. Here, we explain this issue in further detail.

In TDRL, the critical assumption is that all relevant information to predict the future is encoded by the current state of the environment. This is the Markov assumption. While it is possible to bake in the entire history into the definition of the current state, such an approach is not practically useful. Hence, the current state in cue-reward tasks is typically considered to be measuring the time since the cue onset, with this state information sufficient to determine the future states and reward (can also be partially observable). This means that every stimulus that can in principle predict a reward should evoke its own delay states, since in the absence of such states, it is not possible to assign value to those states and learn the stimulus-reward association. On the other hand, retrospective causal learning assumes that recent history is maintained in a timeline. A set of eligibility traces for each stimulus provide such a timeline (e.g., (*65*)).

Using these representations, an approach to reduce the number of stimuli under consideration for associative learning (either prospectively or retrospectively) is to filter them by their salience (e.g., only stimuli crossing a minimum salience threshold will be considered for associations with reward). Performed prospectively, any stimulus below this threshold will never be learned about, since no state space measuring time since its onset will be constructed (as the stimulus did not cross the minimum salience threshold). Thus, the threshold for salience is critical: set it too low and you have too many stimuli to learn about but set it too high and you may miss important associations. On the other hand, since retrospective learning requires a timeline of recent experience in the brain, it is possible to adaptively set the salience threshold or to speed up learning by evaluating previous stimuli ordered by their salience.

To illustrate this difference, let us consider the conditioned taste aversion result wherein saccharin consumption followed 24 hours later by the induction of illness resulted in the animal avoiding saccharin thereafter (*90*). Such learning is extremely hard with prospective TDRL since all stimuli with the salience of saccharin need to be considered for the potential possibility that they may be paired with a meaningful outcome such as illness at least up to 24 hours later. This requires a state space measuring time since stimulus for at least 24 hours that does not get terminated by any of the countless other potential outcomes during those 24 hours. This is highly implausible. On the other hand, looking retrospectively from illness, the animal can rank order recent stimuli based on their salience and rapidly assign credit for the illness to saccharin. Thus, the possibility of rank ordering previous stimuli by their salience allows a more rapid retrospective associative learning.

There is a further question about whether the memory requirements for TDRL using standard state spaces is the same as that for retrospective ANCCR learning. There are two ways that TD state spaces are constructed, and neither is equivalent to the retrospective approach. The most common approach in TD is to *assume that the animal knows what a "trial" is*. In this case, one can construct a Rescorla Wagner state space in which there is one state for the whole "trial" following a stimulus. It is true in this case that there only needs to be one associative weight and one eligibility trace for each candidate stimulus, like that for ANCCR. However, it is problematic to simply assume that animals know what a "trial" is (e.g., see **Fig S5**). The more common approach is that TD models measure time following each stimulus using a cascade of states (CSC, microstimulus, etc.). Here, there are *many* states, associative weights, and eligibility traces per stimulus. Worse, because there is no *a priori* knowledge of the cue-reward delay (or whether there is even a reward that follows a given cue) (see (*21*) for a longer discussion), the number of states should in principle be infinity per stimulus (e.g., animals learn

conditioned taste aversion over even 24 hours). Of course, contemporary neuroscience TDRL models do not run into this problem because they *assume* that animals *a priori* know what a "trial" is and thus, only need to consider T/Δ states per stimulus, where T is the cue-reward delay and Δ is the time resolution for a state (slightly more complex for microstimulus, but the same argument applies). However, aside from the problematic "trial" assumption, this state space still requires the storage of *many* associative weights, and eligibility traces per stimulus, and the memory requirement grows linearly with the cue-reward delay. Thus, the standard prospective TDRL approach and our retrospective ANCCR approach do differ in memory requirement.

Note 8: Responses to punishment:

Our algorithm works similarly for punishments, which are also meaningful causal targets. ANCCR should be multiplied by some measure of intensity of a reward/punishment (Methods). This may either be its incentive value or the absolute magnitude of the incentive value (often called salience). When ANCCR is multiplied by the absolute magnitude of the incentive value, it signals whether a cause predicts any salient meaningful causal target. Instead, when it is multiplied by the signed incentive value, it signals the number of additional (discounted) positively or negatively valenced meaningful causal targets predicted by a given cause. This latter quantity may be used as a signal for incentive salience (i.e., the guidance of approach/avoidance). Hence, if one assumes that different dopamine systems (e.g., those projecting to nucleus accumbens core versus medial shell) signal ANCCR multiplied by the unsigned or signed incentive value of a causal target, these systems may differentially contribute to either learning or motivation.

Note 9: Absence of a backpropagating bump of dopamine activity:

In a simple cue-reward trace conditioning, we observed no backpropagating bump of dopamine activity from the moment immediately preceding reward to cue onset (**Fig 6**). This contrasts with an apparent backpropagation in a recent publication (*50*), which was the first among many attempts to clearly find evidence for such an effect during initial learning of a cue-reward association. The authors argued that they were able to observe such a backpropagating bump due to the measurement of dopamine release in ventral striatum using fiber photometry during initial learning. Since we employed the same approach, we believe that the discrepancy may be due to another reason. Specifically, the previous publication used an olfactory stimulus for trace conditioning whereas we used an auditory stimulus. Olfactory stimuli do not have a clearly defined offset like auditory stimuli. This is because odors can last in the nasal cavity of an animal for longer than the intended duration of presentation. Further, odor detection is an active process and requires respiration, thereby implying that there is also not a well-controlled onset of odor detection. Hence, an alternative explanation for the apparent backpropagating bump is that animals become better at detecting odors over learning, such that odor detection becomes progressively closer to the true onset of presentation. If dopamine simply responds to the moment of detection of the odor, it will exhibit an apparent backpropagation within a trial over learning, as was observed.

Note 10: Magnitude of inhibition of dopamine release in Fig 6:

In the optogenetic experiment in Fig 6, we observed a linearly decreasing dLight response during the 1.5 s inhibition of VTA dopamine cell bodies. Though the apparent reduction in dopamine is gradual, we think that this merely reflects binding of the remaining dopamine in NAcc and that there is an abrupt reduction of NAcc dopamine release from the laser onset. Hence, for simulations of the RPE model, we used the eventual dopamine level relative to baseline at 1.5 s following inhibition onset. This was ~0.6 times the magnitude of the reward response. Even if the negative RPE was assumed to be a linearly decreasing function until 1.5 s, TDRL predicts a negative cue1 response (**Fig S12C**).

There are two caveats related to the ANCCR prediction of intact cue1 learning in this experiment. The first caveat results from a side effect of the strong inhibition, viz., that the reward response relative to pre-cue baseline in the experimental group was as low as the eventual asymptotic reward response after learning in the control group (**Fig 6**). Thus, ANCCR predicted slower behavioral learning in this experiment due to the smaller reward response. Second, due to inhibition of cue2 response, ANCCR predicts that the learning of the cue1→cue2 association will be disrupted. This is because cue2 no longer becomes a meaningful causal target due to the inhibition. Accordingly, ANCCR predicts that cue2 will be treated as an "unexplained" cause of reward (instead of as a cause already explained by cue1). Thus, the resultant overexpectation of reward will result in a lower eventual cue1 response when normalized by reward response on day 1 of conditioning (**Fig 6**). Consistent with ANCCR, we observed that every experimental animal learned the task, albeit slower than control animals (**Fig 6I**), and that mesolimbic dopamine acquired positive responses to cue1 in all experimental animals, albeit at a lower asymptote relative to the control group (**Fig 6I**).

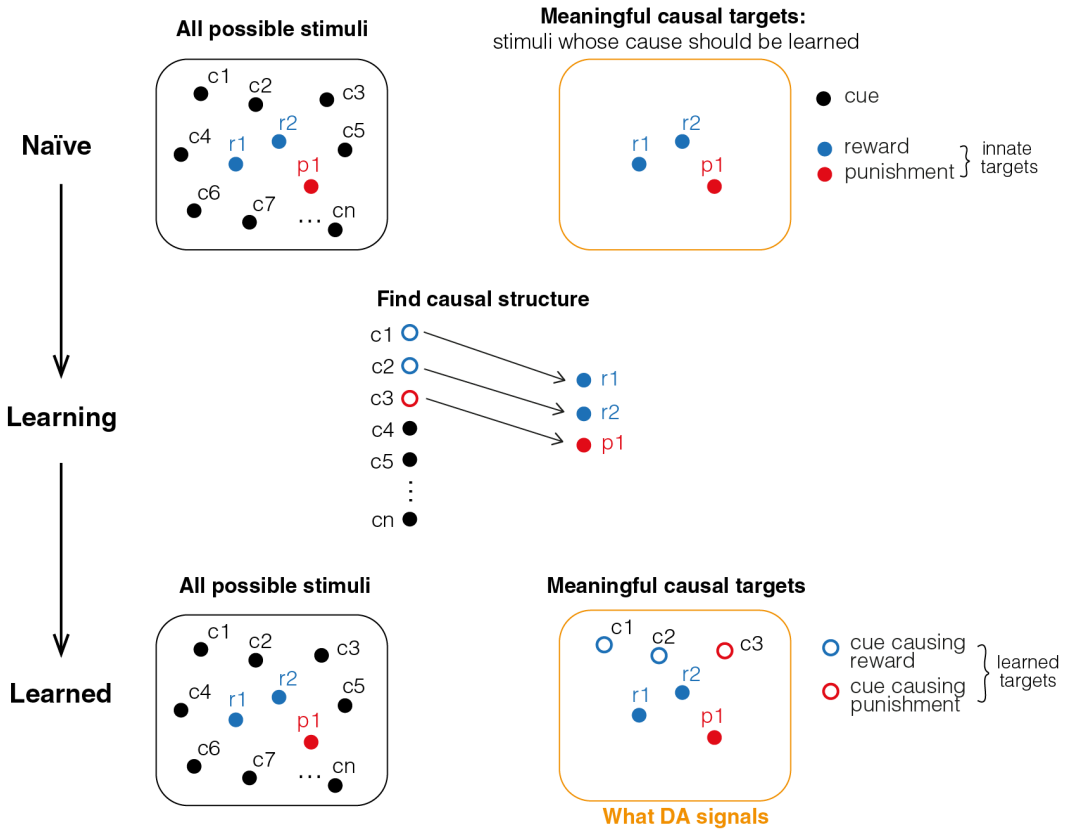


Fig. S1.

Simplified schematic of learning and mesolimbic dopamine function. In a naïve animal, some stimuli are innately regarded as meaningful causal targets. After learning the appropriate causal structure between neutral stimuli and meaningful causal targets, the neutral stimuli that predict meaningful causal targets themselves become meaningful causal targets. We hypothesize that mesolimbic dopamine conveys the learned meaningfulness of a current stimulus. Please note that we hypothesize that the primary purpose of the mesolimbic dopamine system is to infer how meaningful a given stimulus (including rewards) is after learning. Thus, though a reward is an innately meaningful stimulus, its dopamine response will change with learning (**Fig 3**).

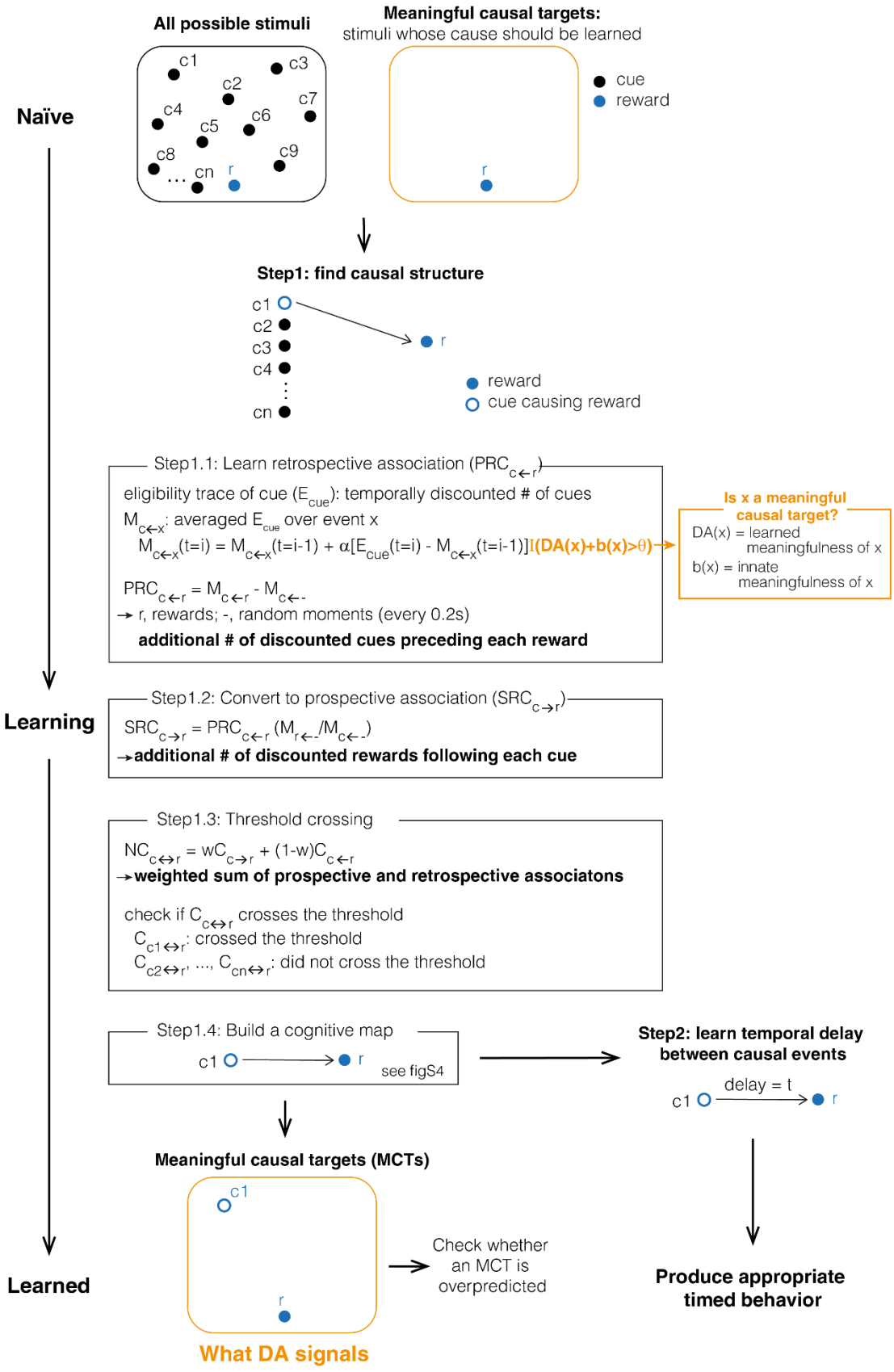


Fig. S2.

Extended schematic showing the steps for learning. Broadly, we hypothesize that there are at least two steps prior to behavioral learning: learning a causal model of associations, and the learning of the temporal delays between causal events. See Methods for details of each step. The rationale for Step 1.2 is shown in **Fig S3** and for Step 1.4 is shown in **Fig S4**. The

threshold crossing in Step 1.3 implies that the net contingency signal will accumulate gradually during learning, but the downstream behavioral learning will be abrupt after threshold crossing and temporal learning. Together, these steps allow the formation of a causal cognitive map detailing the predictors of meaningful causal targets. At this stage, this map does not contain a representation of the temporal delay between these causal events (see **Fig S6** for why such learning is non-trivial). Here, we do not directly propose an algorithm for learning the relevant temporal delay. Once a causal cognitive map with temporal delays is learned, the animal can produce an appropriate timed behavior following the presentation of a cue that predicts reward.

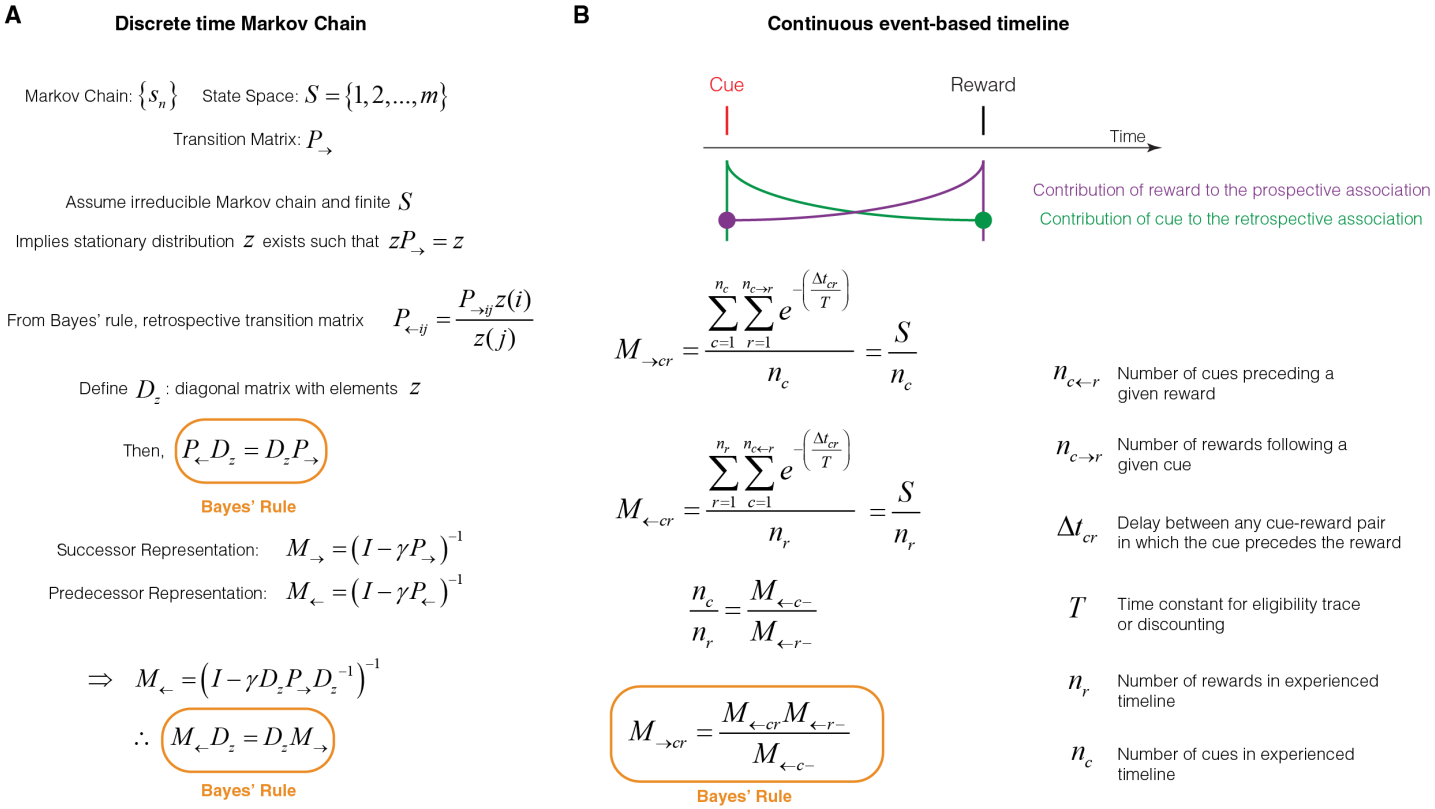


Fig. S3.

Bayesian relationship between prospective and retrospective representations. A. Derivation of the Bayes relationship between successor and predecessor representation in an irreducible discrete time Markov chain with a finite state space (*20*). **B.** Extension of this approach to a continuous event-based timeline. Intuitively, for any given cue-reward pair in a timeline of experience, the contribution of the reward to a prospective successor representation estimate at the cue is equal to the contribution of the cue to a retrospective predecessor representation estimate at the reward. Thus, the only difference between the successor and predecessor representation between cue and reward is that the former is averaged over the cue occurrences while the latter is averaged over reward occurrences. Therefore, if the base rate of cues and rewards are known in an environment, the successor representation can be calculated from the predecessor representation using Bayes' rule (Supplementary Note 1).

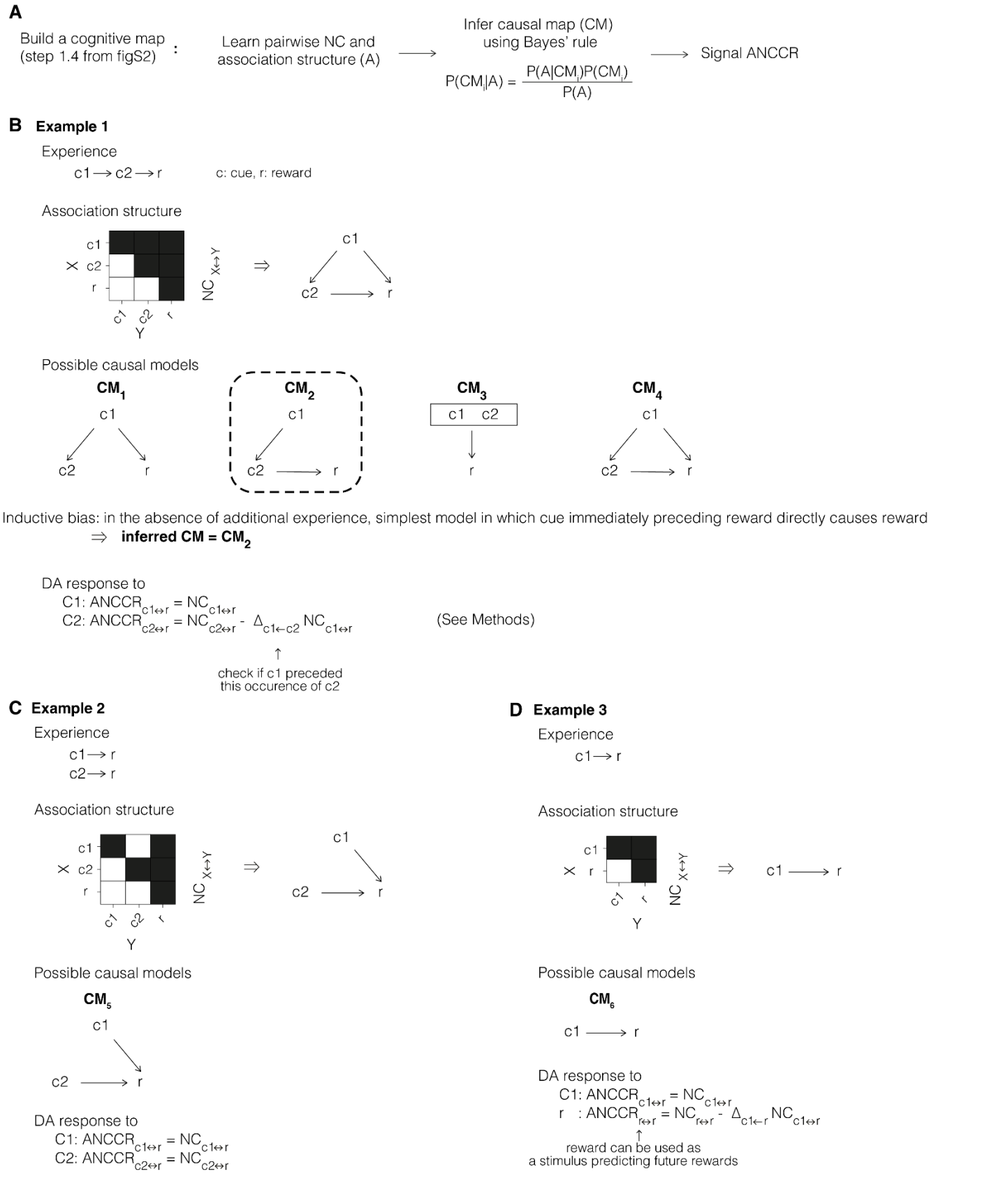


Fig. S4.

Causal cognitive map inference and calculation of ANCCR. A. Once pairwise net contingencies are learned between events and meaningful causal targets (using Equation (13) in Methods), the animal can use this net association structure A

(joint distribution of all net contingencies) to infer appropriate causal models using Bayes' rule, following which the ANCCR for all event types (i.e., states) can be calculated. **B.** An example illustrating ANCCR. If an animal repeatedly receives the sequence cue1 followed by cue2 followed by reward, and no other experience, the animal can eventually learn the displayed pairwise association structure. Four possible causal models are compatible with this association structure. For simplicity, we will assume that the association structure is binary (i.e., does a given causal arrow exist or not following thresholding?). Causal model CM_3 assumes that cue1 followed by cue2 is treated as a single compound stimulus (denoted by the rectangle). We assume that animals use an inductive bias whereby in the absence of additional experience, they assume the simplest model in which the last cue preceding reward causes the reward. This inductive bias implies that animals would assume that the individual presentation of cue2 will result in reward, which rules out the first and third causal models. Hence, given the experience of cue1 followed by cue2 followed by reward, we assume that the animals infer the underlying causal model to be the simplest among CM_2 or CM_4, i.e., CM_2. ANCCR to any cue measures whether it must be treated as a meaningful causal target. Here, cue2's occurrence is predicted by cue1. Hence, cue1 is a more important meaningful causal target than cue2. We use this intuition to define the ANCCR for cue2 as the pairwise net contingency between cue2 and reward subtracted by the pairwise contingency of cue1 and reward if cue1 precedes the current occurrence of cue2. This instantaneous precedence can be measured by a non-cumulative eligibility trace of cue1 at cue2. Since there is no incoming arrow into cue1 in the causal model (i.e., no backdoor path (*81*)), its ANCCR is its pairwise association with the reward. **C.** Similar calculation for an environment in which the experience involves cue1 and cue2 independently predicting reward. Here, since there is no incoming causal arrow into cue2, its ANCCR does not include the negative corrective term. **D.** For a simple cue-reward conditioning, the reward can itself be a predictor of reward. Thus, its ANCCR will involve its pairwise net contingency with itself (which depends on the temporal schedule of rewards in the environment) minus the corrective term resulting from the incoming causal arrow to the reward (from the cue). Hence, the reward response will reduce once a cue is learned as a cause of the reward. As a corollary, if rewards become subsequently omitted, an omission ANCCR response will appear only if the omission of reward is itself inferred as a state. Such inference is reasonable when reward probability is reduced from 100%, since the animal needs to infer that the lack of a reward following the cue on a given trial should imply that the animal should suppress reward seeking. However, if the delay to a reward is increased permanently, the omission of the reward never gets treated as its own state, since the animal can simply update its estimate of the delay to reward. This is tested in Experiment 3, Test 5 (**Fig 4**).

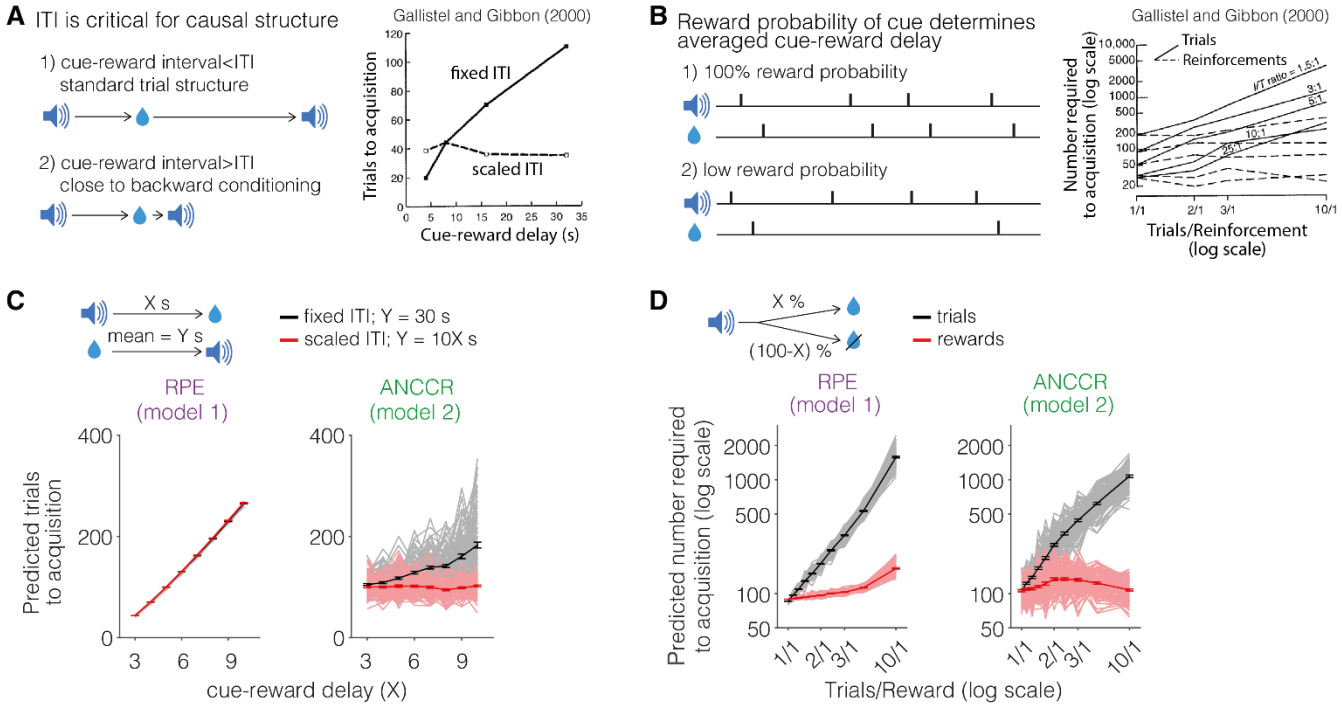


Fig. S5.

Retrospective causal learning explains timescale invariance observed in animal learning. A. We have previously discussed why a general causal learning algorithm is likely invariant to a scaling of temporal intervals (*21*). Here, we give an intuition for this argument by showing that the intertrial interval (ITI) in cue-reward learning is critical for the causal structure. Imagine two experiments with the same cue-reward delay. Typical TDRL algorithms only model the cue-reward delay under the assumption that it is the only relevant "trial period". However, imagine two extremes: one with a very long ITI and another with a very short ITI. In the long ITI case, the delay from cue to reward is shorter than the delay from a reward to the next cue. However, in the short ITI case, the delay from reward to next cue is shorter and hence, the task appears like "backward conditioning", wherein the "trial" can now be thought of as the reward to cue interval. This shows that both the "trial period" and the structure of the task critically depend on the cue-reward delay *as well as the ITI*. The right panel contains actual behavioral results reproduced from (*82*) showing that when the cue-reward delay is increased, the number of trials taken by animals until acquisition increases only if the ITI is held fixed, and not if the ITI is correspondingly scaled. This is a demonstration of timescale invariance of initial learning since scaling the time delays equally (both cue-reward delay and ITI) does not change the number of trials required until acquisition. **B.** A similar invariance is manifested when the probability of reward is reduced while maintaining a fixed cue-reward delay and ITI. A reduction in reward probability can be equivalently thought of as an increase in the average delay from a cue to the next reward. Thus, timescales and probabilities are closely tied to one another. The right reproduced plot shows that when the probability is reduced (i.e., number of trials per reinforcement is increased), the number of trials until acquisition in real behavioral data depends only on the number of reinforcements. **C.** Simulations demonstrating that TDRL cannot explain the timescale invariance, but that retrospective causal learning explains these results (Methods). **D.** Simulations demonstrating that TDRL cannot explain the invariance of rewards to acquisition with respect to probability of reinforcement, but that retrospective causal learning does (Methods).

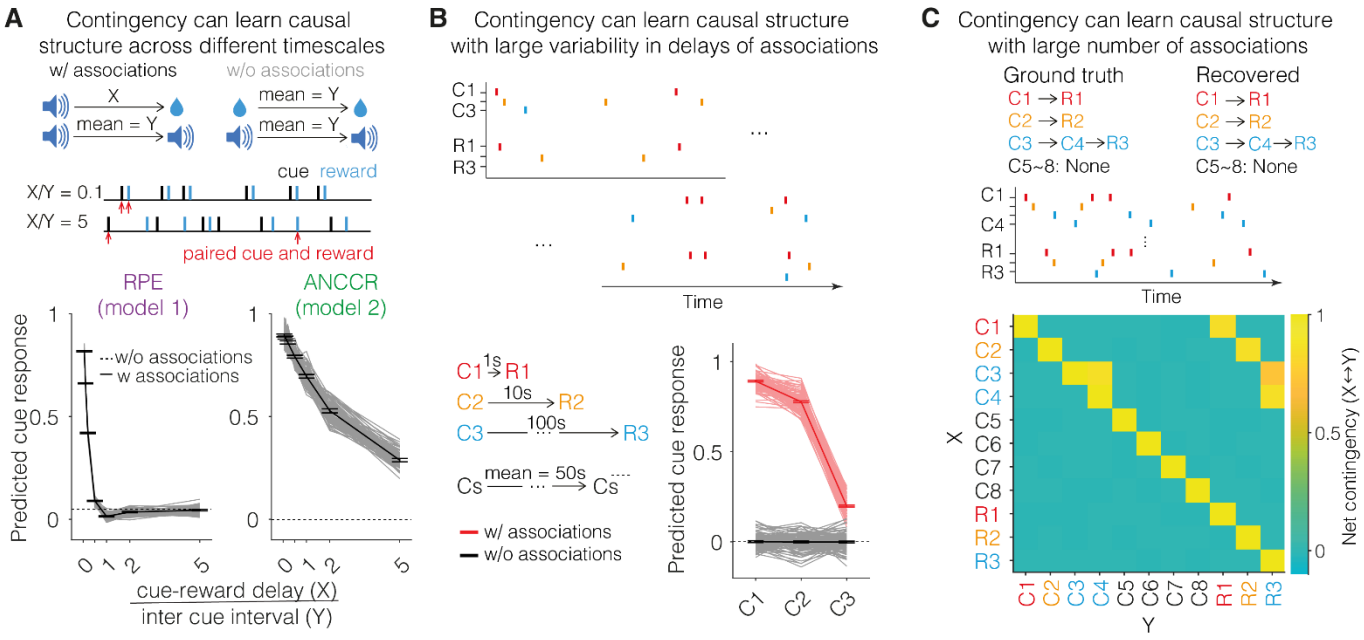


Fig. S6.

Contingency learning recovers true causal structure without loss of generality across timescales A. Here, we show simulations of learning of what is often considered the simplest associative learning problem: a cue predicts a reward with 100% probability at a fixed delay (denoted by X). We remove the simplifying assumption that the next cue following a "trial" comes only after the current reward. Specifically, we assume that the cues themselves arrive with an exponential inter cue interval (mean=Y) (labeled w/associations in the schematic). As a control condition, we consider the presentation of rewards and cues at the same rates but with no temporal relationship between each other (labeled w/o associations in the schematic). We systematically varied the X/Y ratio and tested if TDRL and causal contingency learning learn the cue-reward association. When X/Y is very small, other events are unlikely between any given pair of cue and reward. However, when X/Y is large, other cue/reward events are common between any given pair of cue and reward (an example timeline is shown with red arrows pointing to paired cue and reward). TDRL learns an RPE for the cue that is often lower than that when there is no association, implying that this algorithm cannot learn that every reward is predictable by a previous cue (in fact may even learn that a cue decreases value). However, across all ratios shown, contingency learning signals a significantly positive ANCCR for the cue, implying that the algorithm learns the cue-reward association without loss of generality. Though not explicitly shown here, this simulation shows that learning the temporal delay between a paired cue and reward is very challenging despite learning a cue-reward association. For instance, in the X/Y=5 case, simply timing the interval between one cue and the next reward will not allow the learning of the cue-reward delay. The only systematic solution to this problem is to have a pool of eligibility trace time constants (*64–66*, *83*), a complexity that we do not consider here. **B.** In a more complex world, we assumed that there are three simultaneously occurring associations between three separate pairs of cue and reward, but each with vastly different delays. An example timeline shows that every cue1 is followed by reward1 quickly, cue2 followed by reward2 at a longer delay with some intermediate events, and cue3 followed by reward3 with many intermediate events. Since TDRL could not even learn the simpler problem considered in **A**, we did not simulate TDRL in this task. Even with three timescales of associations separated by two orders of magnitude, our algorithm learns the correct relationships. This was true despite there being only one eligibility trace time constant for all stimuli (Methods). **C.** When the number of causal associations in the world increases with many cues being "distractors" without any associations with other events (cues 5-8), the learned net contingencies perfectly recovered the underlying causal structure (Methods). The bottom matrix shows the recovered net contingency structure from the simulation and learns all the correct relationships. Please note that for this simulation, our intent was to show that net contingency reflects causal structure and to show the entire matrix, we assumed that every event was a meaningful causal target.

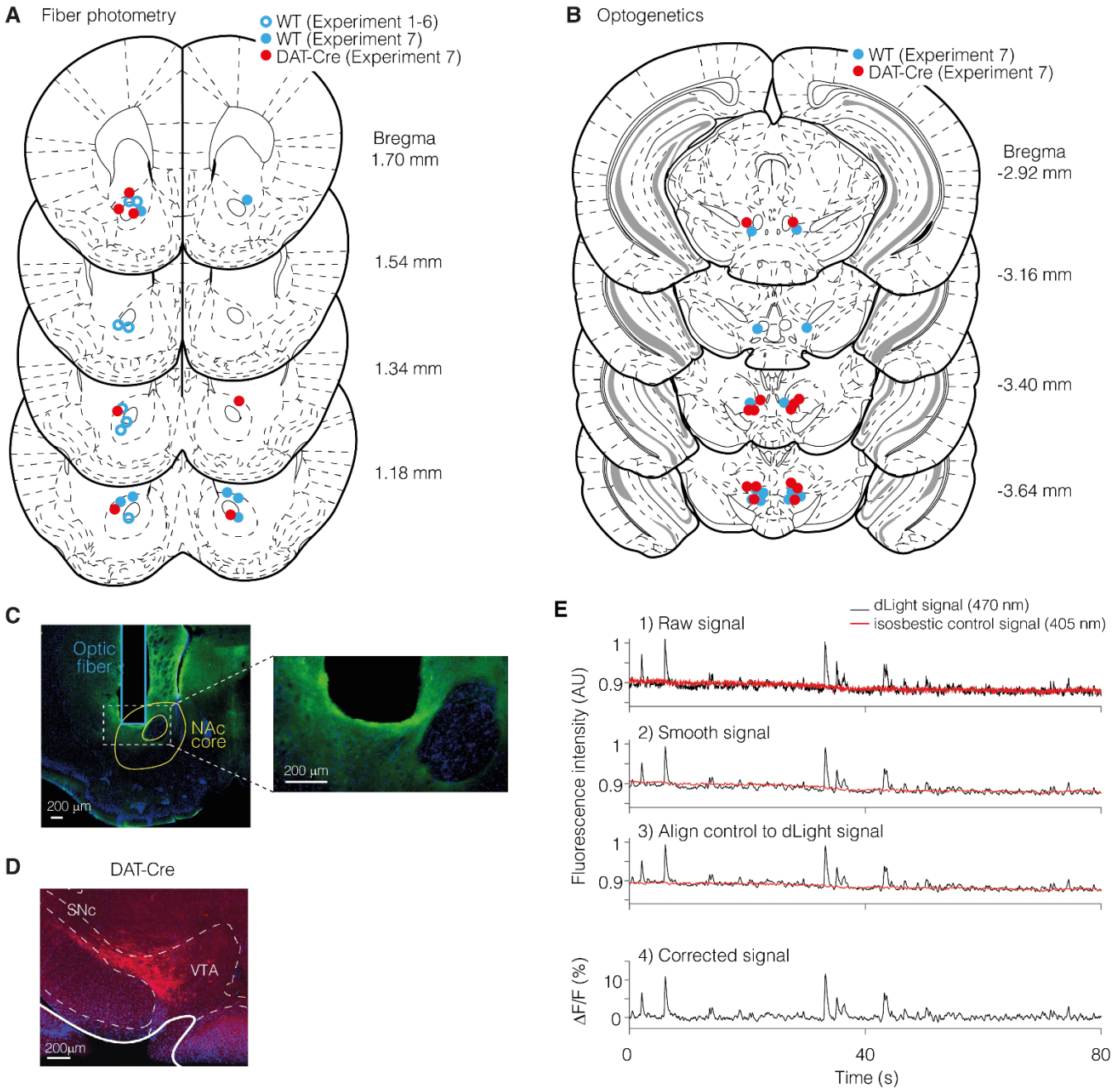


Fig. S7.

Fiber photometry to measure dopamine release using dLight 1.3b. A, B. Reconstructed centers of the optical fiber for fiber photometry (**A**) and optogenetics (**B**) for all animals. In all cases, the center of the fiber was in NAcc (**A**) and VTA (**B**). **C, D.** Example histology of one animal showing dLight (**C**) and stGtACR2-FusionRed (**D**) expressions. **E.** Fiber photometry analysis approach to obtain a baseline corrected signal (Methods).

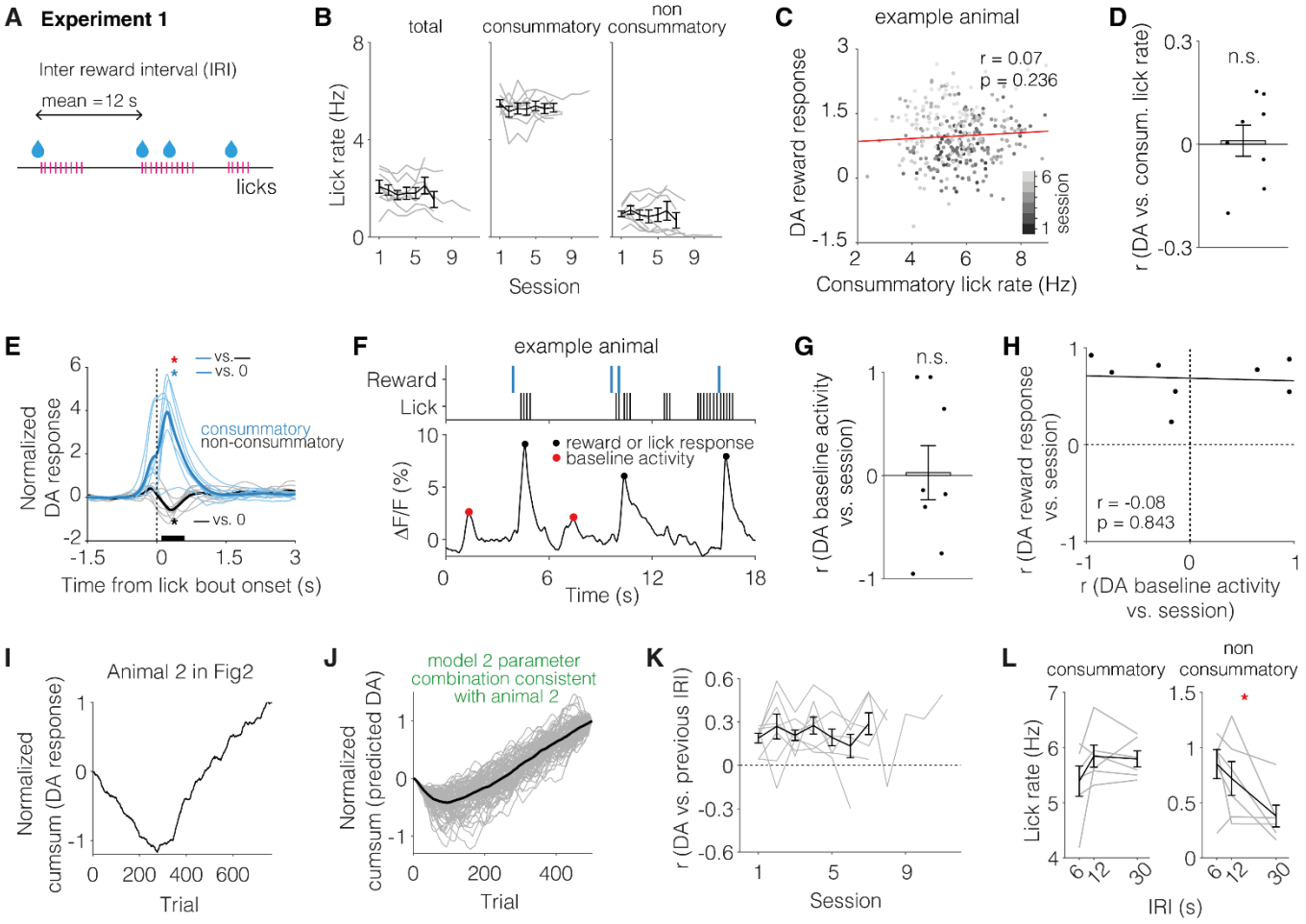


Fig. S8.

Dopamine responses during random rewards. A. Schematic of the random rewards experiment (replicated from **Fig 3**). **B.** Average lick rate across sessions across all animals for consummatory (related to sucrose consumption) and non-consummatory licks (unrelated to sucrose consumption) (see Methods). The lick rates generally remained constant across repeated sessions. See **Supplementary Table 1** for all statistical quantification of all statistical results. **C-D.** The consummatory lick rate is not correlated with dopamine reward response. **E.** The average lick bout onset response was positive for consummatory lick bouts (during sucrose consumption) but was slightly negative for dry lick bout onsets away from reward consumption. Dopamine response was z-normalized by the response during [-1.5 -0.5] s from lick bout onset. This shows that the increase in reward response with repeated experience cannot be due to a non-specific increase in a positive lick bout onset response. This is because the lick bout onset response in the absence of reward is negative. **F.** Example animal showing the identification of baseline dopamine activity unrelated to rewards or lick bouts. Such activity is useful to evaluate whether the increase in dopamine response across sessions is specific to sucrose or is also true of baseline responses. **G.** There was no change in the baseline dopamine activity across sessions. **H.** Dopamine baseline activity change across sessions is uncorrelated with dopamine reward response change across sessions. This shows that the increase in sucrose responses across experience is unlikely to be explained by restraint stress (Supplementary Note 3). **I.** Reproduction of the dopamine reward response from animal 2 in **Fig 3**. **J.** Simulation parameters that are consistent with the pattern observed in animal 2, with dopamine response to unpredicted sucrose starting out negative but then switching to reach a positive asymptote (i.e., constant positive slope in the cumsum plot). **K.** The correlation between dopamine response to a reward and the previous IRI remains positive across sessions. **L.** A separate group of 6 mice was trained on the random rewards experiment with three different IRIs, 6, 12, and 30 s (one per session) to test whether their non-consummatory lick rates reflected the expected reward rate in a session. Each IRI condition was trained for two consecutive sessions. The order of the IRI condition was randomly assigned for each animal. The non-consummatory exploratory licks made by the animals depended on the average IRI for a given session. Thus, animals learn the average reward rate of an environment within at most two sessions.

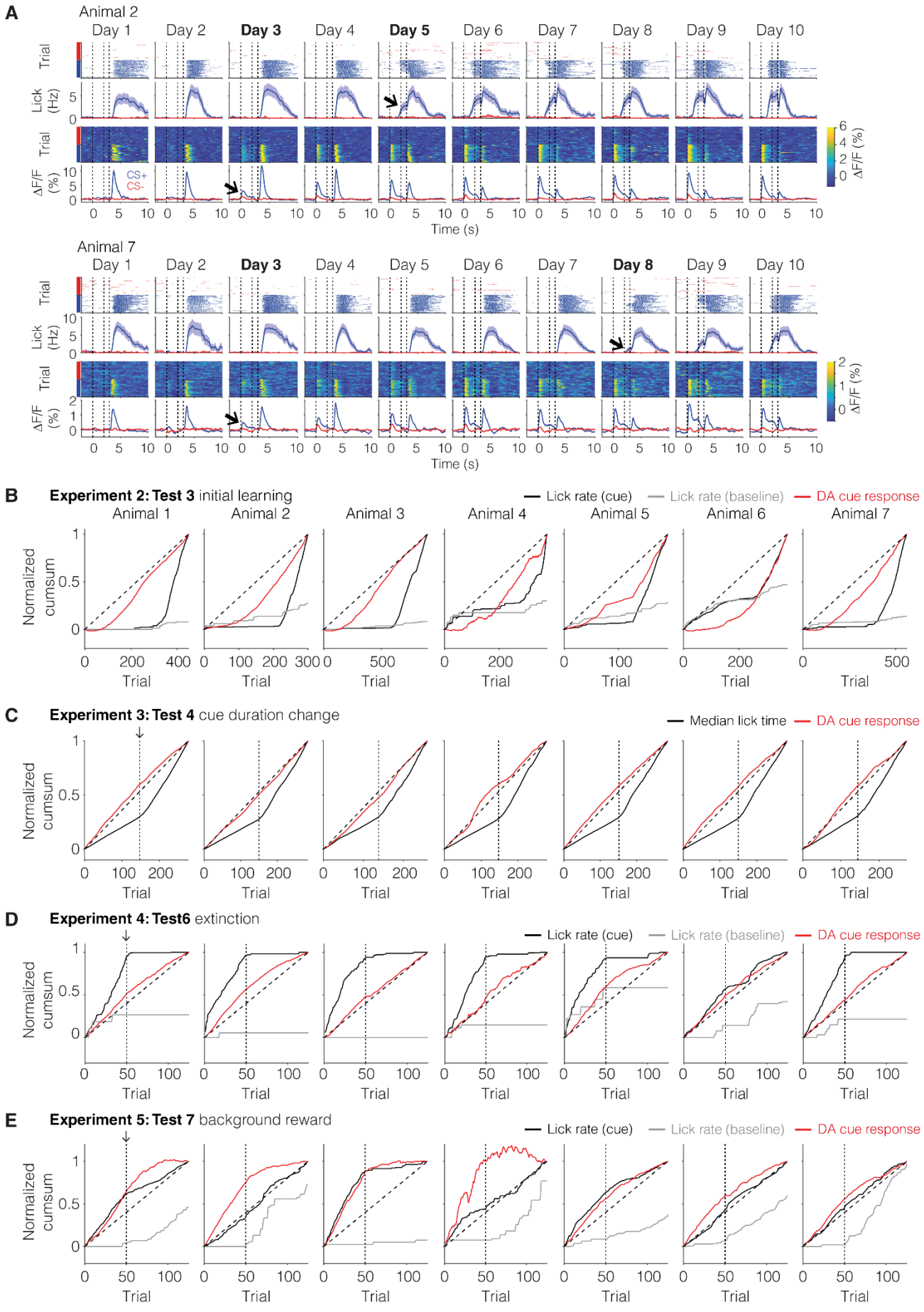

A
Animal 2
Day 1
Day 2
Day 3
Day 4
Day 5
Day 6
Day 7
Day 8
Day 9
Day 10
Trial
Lick (Hz)
ΔF/F (%)
CS+
CS-
Time (s)
Animal 7
B
Experiment 2: Test 3 initial learning
Lick rate (cue)
Lick rate (baseline)
DA cue response
Animal 1
Animal 2
Animal 3
Animal 4
Animal 5
Animal 6
Animal 7
Normalized cumsum
Trial
C
Experiment 3: Test 4 cue duration change
Median lick time
D
Experiment 4: Test6 extinction
E
Experiment 5: Test 7 background reward

Fig. S9.

Individual animal data during cue-reward learning. A. Similar plots as **Fig 4C** from two more example animals. **B-D.** The raw animal-by-animal data for baseline lick rate, cue-period lick rate, and baseline subtracted dopamine response for all animals corresponding to experiments 2-5. When unpredicted rewards were delivered during the ITI in experiment 5, the baseline lick rate increased without a corresponding increase in the cue period lick rate.

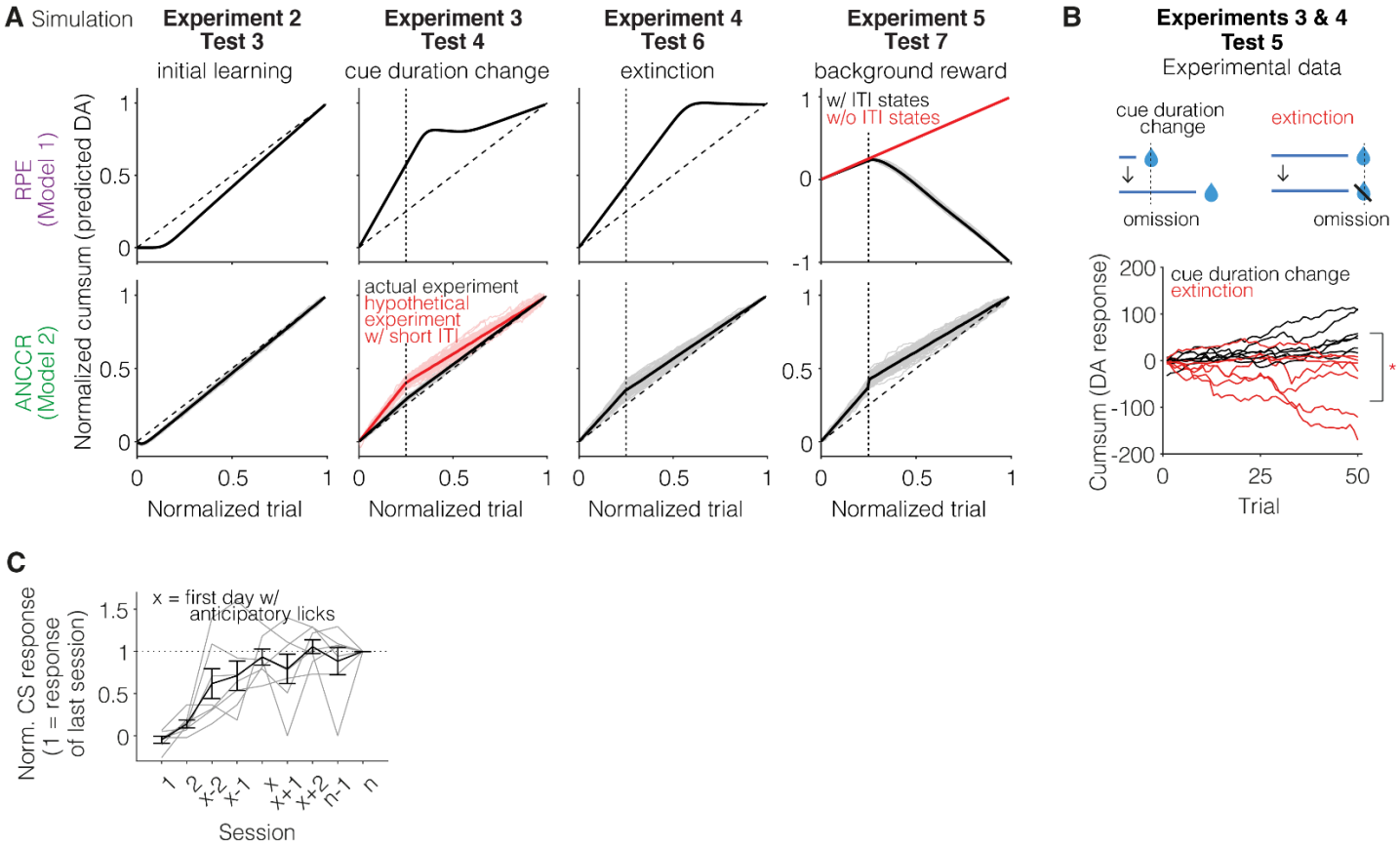


Fig. S10.

A. TDRL and causal contingency models show different dynamics for the experiments in **Fig 4**. Observed results are only consistent with the causal contingency model. **B.** Omission response is not negative with respect to pre-cue baseline for any animal after cue duration change but was negative during extinction. **C.** Observed dopamine cue responses when animals showed behavioral acquisition (marked as session x). Animals were trained until they show asymptotic behavior. Session n is the final session. Dopamine cue response reaches 93% of its response on the final session (~100% if normalized relative to average cue response on session n and n-1). The significance of this observation is discussed in Supplementary Note 5.

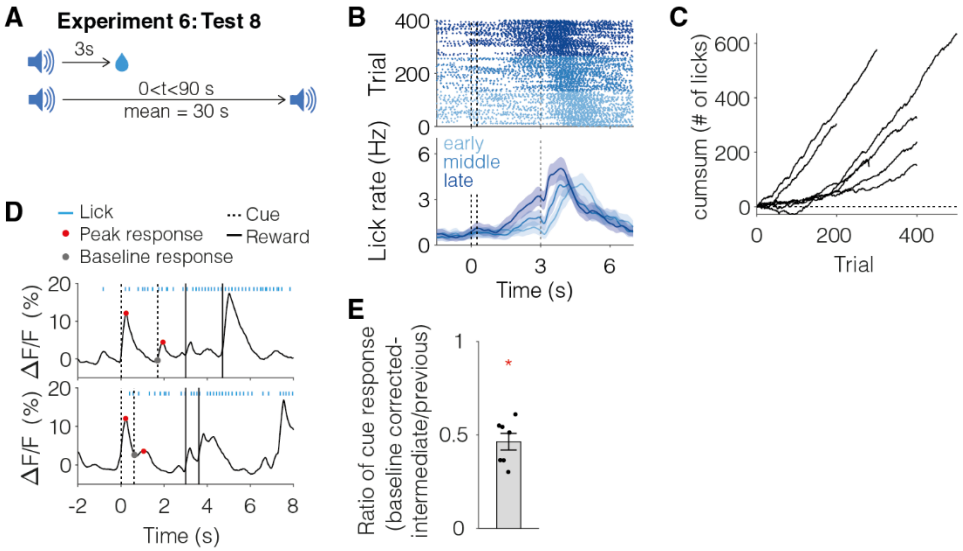


Fig. S11.

A. Schematic of the "trial-less" cue-reward task (reproduced from **Fig 5A**). **B-C.** Animals rapidly learn the "trial-less" cue-reward task even though the cue duration is short (250 ms), trace interval is long (3 s), and there are intermediate cues. Cumulative sum of anticipatory lick during cue-reward delay are shown in **C**. **D-E.** Here, we measure baseline as the dopamine fluorescence immediately prior to the intermediate cue delivery, as opposed to the baseline prior to the first cue in **Fig 5C, D**. Cue response was defined as baseline subtracted peak response. The intermediate cue response was still positive. In other words, the positive intermediate cue response is not an artifact of subtracting the pre-cue baseline.

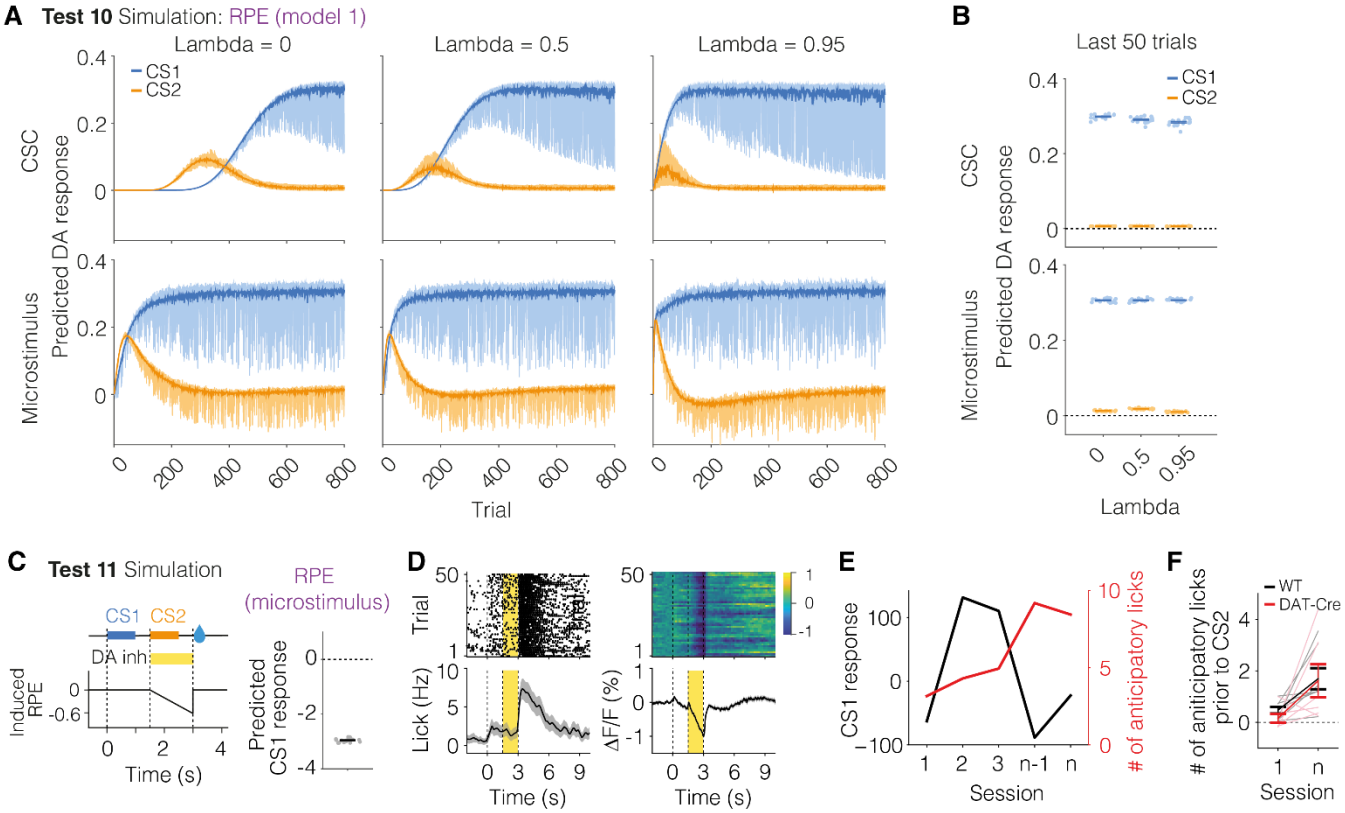


Fig. S12.

TDRL RPE predictions for optogenetic inhibition during sequential conditioning. A. Simulation results from CSC and microstimulus TDRL models for various lambda. **B.** Asymptotic responses for CS2 were close to zero for both models. **C.** Simulation results for TDRL assuming a negatively ramping RPE. **D.** Outlier animal in which inhibition was so strong that the reward response was negative with respect to pre-CS1 baseline on day 1 of conditioning. Due to this, we excluded this animal from analysis. **E.** Evolution of CS1 dopamine response and licking from this animal. Though CS1 response and behavior increased quickly in this animal, CS1 response became unstable after learning. **F.** Quantification of pre-CS2 anticipatory licks showing reward-seeking prior to CS2 (onset of dopamine neuron inhibition).

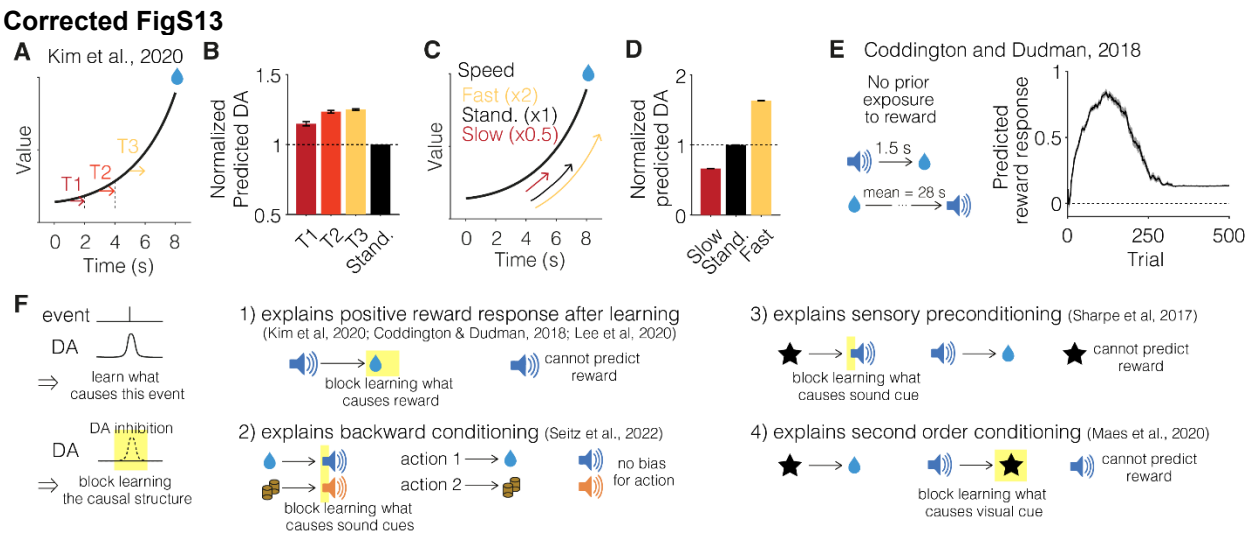


Fig. S13.
Causal contingency model explains multiple recent findings related to DA responses. A-D. Simulations showing that recent observations of dopamine ramping that were argued to demonstrate RPE signaling (*58*) are also consistent with ANCCR signaling (Methods). ANCCR signaling not only ramps up when a series of cues predicts reward, but also captures observed changes in dopamine activity when a subject was teleported to a moment closer to reward (**A-B**) or the speed of cue presentation changed (**C-D**). **E.** If cue-reward learning is performed without prior exposure to reward in the experimental context, the predicted dopamine reward response increases initially prior to reaching a smaller positive asymptote. This was previously observed and is inconsistent with RPE signaling (*54*). **F.** A schematic showing the hypothesis that mesolimbic dopamine signals whether a currently experienced stimulus is a meaningful causal target, thereby promoting the learning of its cause. Dopamine inhibition during the event prevents causal learning in this framework. This conceptually explains positive reward responses for fully predicted delayed rewards after learning, the role of dopamine in backward conditioning, sensory preconditioning, and second-order conditioning. Please note that since the original experiments were not projection-specific (e.g., just modulating dopamine projections to NAcc), we assume here that these experiments also manipulate dopamine systems that signal that highly salient cues are meaningful even prior to any association with rewards (e.g., cues that cause startle responses).

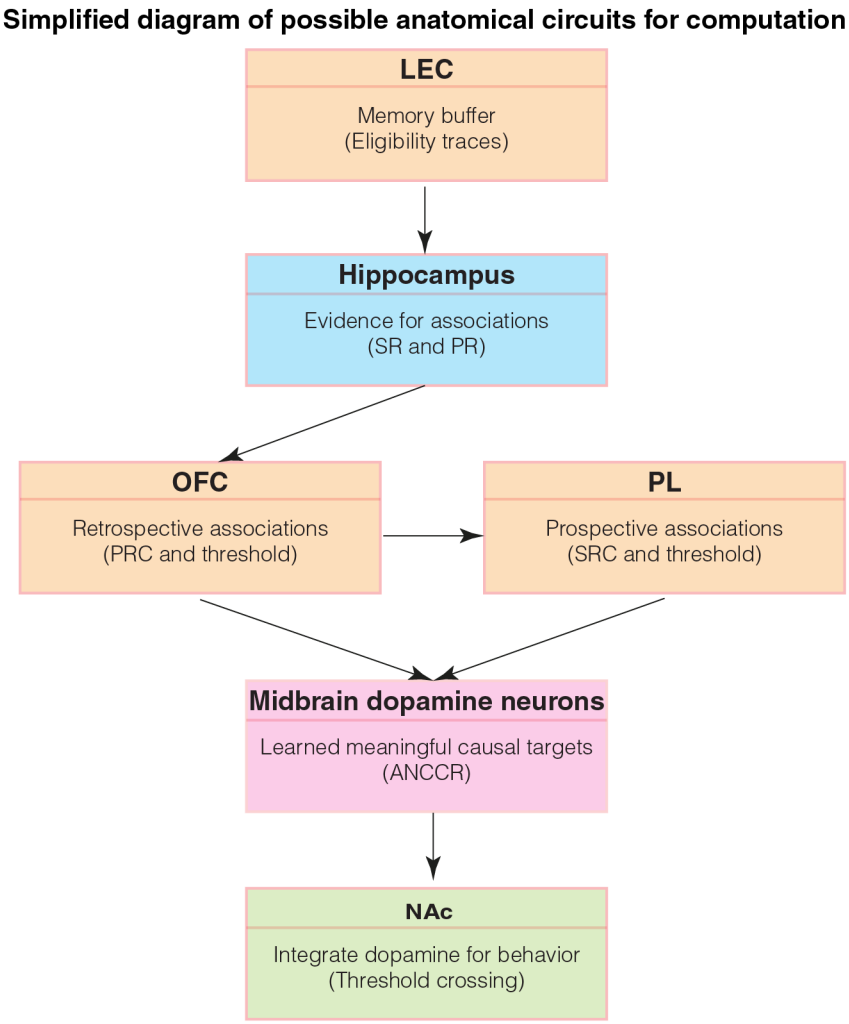


Fig. S14.

Possible anatomical basis for the computations upstream of dopamine neurons. A simplified putative circuit for computation of ANCCR and its influence of behavior. Cortical regions, hippocampus, midbrain, and striatum each have their own color.

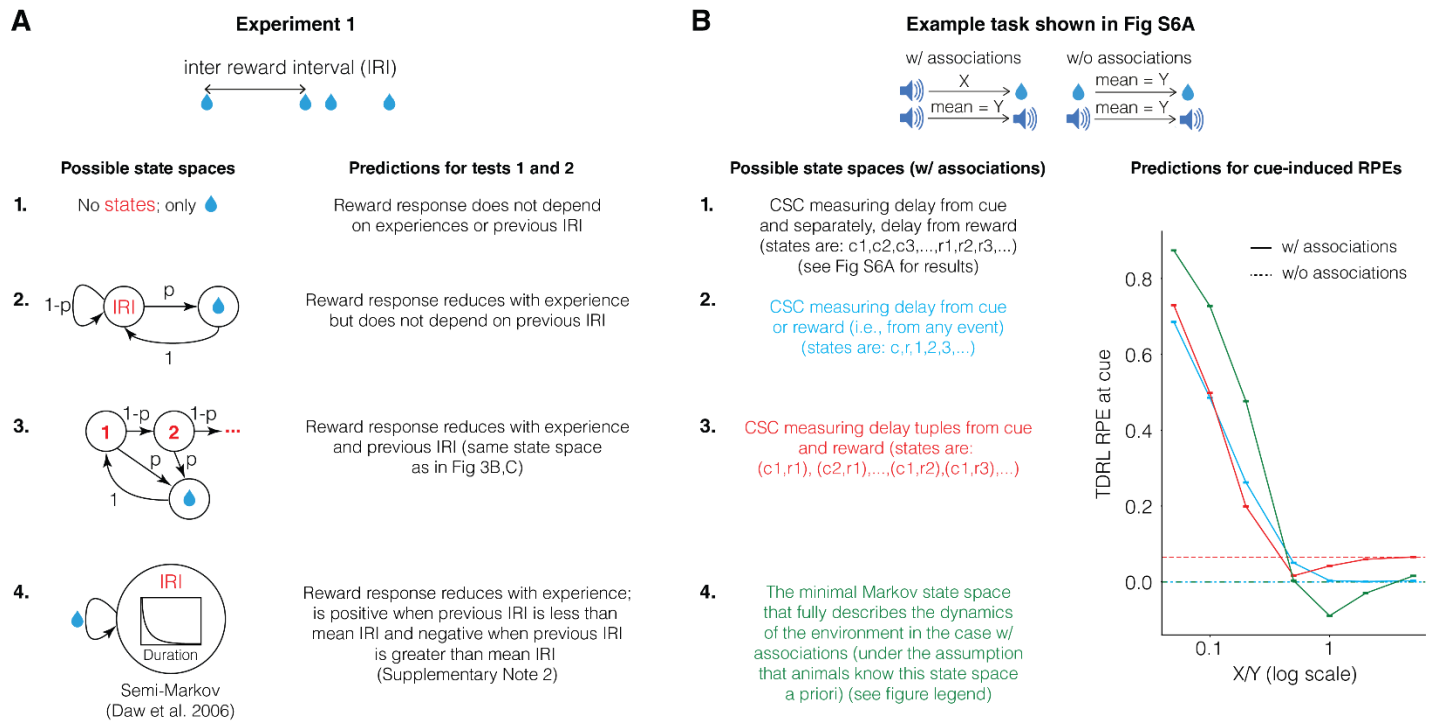


Fig. S15.

Results do not change based on different assumptions of state spaces. Here, we show that our major claims do not depend on the state space assumptions. **A.** Since there is in principle infinite flexibility in assuming state spaces, we first consider the simplest experiment possible (random delivery of rewards). Here, the number of reasonable state space assumptions is limited to the four shown in the graph. The first assumption is that the animal learns no states in the environment, but occasionally gets rewards. This assumption is clearly violated by the fact that animals learn the duration of the IRI (**Fig S8L**), and hence, this state space is not reasonable for these data. State spaces 2 and 3 break the IRI period into either one or multiple states (p is determined by the mean IRI). State space 4 is a semi-Markov one with exponential dwell time in the state. None of these state spaces produce RPEs consistent with the results of Tests 1 and 2. Thus, the results from Tests 1 and 2 violate TDRL for any reasonable state space of the task.

B. Left: Possible state spaces for the associated version of the thought experiment (related to Test 8) from **Fig S6A**. First 3 are variations on CSC states. State space 1: State is "r1", "r2", "r3" etc. for delay after a reward, and "c1", "c2", "c3" etc. for delay after a cue. Both are measured in time bins of 200ms. This is the CSC model assumed in **Fig S6A**. Please note that standard CSC models in neuroscience only measure time within a "trial" and hence, only have the cue states identified here. However, this is obviously a poor description of this task, and hence, we only consider this extension that measures time from cue and reward. State space 2: State is "c" if cue, "r" if reward, else time since last event. State space 3: State is time since last reward and time since last cue, i.e., the combination of delays from the most recent cue and reward. State space 4: This is the minimal state space that can correctly simulate the environment. State is the set of as-yet-unrewarded cue times. Note that the process is not Markov in the CSC state spaces above due to partial observation.

Right: RPE at cue onset, averaged over the last 25% of 2M simulation steps, for the value function learned by TD learning with discount factor 0.95 and learning rate 0.05, for the last 3 state spaces described on the left (see **Fig S6A** for the first), and for different values of reward delay X relative to inter-cue interval Y=12s. Dotted lines are the average RPE values for the non-associated case. We see RPE fails to discriminate the association between cues and rewards unless the reward delay is significantly below the inter-cue interval. This contrasts with ANCCR as shown in **Fig S6A**.

Since the RPE curve for even the full state space (#4) in the associated case unintuitively decreased below the non-associated case, we calculated the exact value function and RPEs for this case analytically. To calculate RPE analytically, we first formally describe the associated state space. We model the associated task as a discrete time Markov chain. For a given reward delay X and average ISI Y, both multiples of a discrete time step dt, the state space is the set of tuples of as-yet-unrewarded cue times $(t_0, \dots, t_n), with\ 0 \leq t_0 < \cdots < t_n < X$ relative to the current time. The state $(t_0, \dots, t_n)$ transitions to $(0, t_0 + dt, \dots, t_n + dt)$ with probability $p = \frac{dt}{Y}$ (new cue arrived) and to $(t_0 + dt, \dots, t_n + dt)$ with probability 1-p (no

new cue). If $t_n + dt = X$, a reward is given, and the time is deleted from the list of unrewarded cue times. This is the minimal state space necessary to simulate the task (#4 above).

The exact value function for this state space can be guessed and verified to be true based on the Bellman equation as

$$V\left(t_1,...,t_n\right) = \sum_{i=1}^{n} \gamma^{\frac{X}{dt}-1-\frac{t_i}{dt}} + \frac{dt}{Y}.\frac{\gamma^{\frac{X}{dt}}}{1-\gamma}$$

Let s' = (t_1+dt, …, t_n+dt, 0) be any cue state. For the previous state, there are only two options: s_1 = (t_1, …, t_n), i.e., no reward in the transition to the cue state, or s_2 = (X-dt, t_1, …, t_n), i.e., a reward coincided with the cue. When computing RPE(s_1, s') = γV(s') - V(s_1), all terms in the sum on the R.H.S. cancel exactly except for the term corresponding to the new cue, which equals $\gamma^{\frac{X}{dt}-1}$ (intuitively, once a cue has arrived, the reward for it is deterministic, thereby being perfectly accounted for in the value function). The same is true when computing RPE(s_2, s') = γV(s') - V(s_2) + 1, where the X-dt term cancels the explicit reward 1 (i.e., reward is fully predicted). The constant term in the end of the expression in the value function also contributes to the RPE as (γ-1) times this constant. We conclude that RPE(s') = $\gamma^{\frac{X}{dt}} + \left(\gamma - 1\right)\left(\frac{dt}{Y}.\frac{\gamma^{\frac{X}{dt}}}{1-\gamma}\right) = \gamma^{\frac{X}{dt}}\left(1-\frac{dt}{Y}\right)$ is a deterministic constant. This is a positive exponentially decaying function of X. This demonstrates that the analytically derived curve is a positive exponentially decaying curve in X, the cue-reward delay, and that the RPE at cue is independent of prior history (assuming that the exact value function has been learned). Hence, the unintuitive shape of the RPE curve in the above simulations likely results from a constant learning rate assumed in the simulations (which introduces bias and does not converge to the true value function). For comparable discounting and eligibility trace time constant, ANCCR and RPE using the full state space might be qualitatively similar at distinguishing the associated and non-associated environments. For instance, for equivalent discounting as that corresponding to the eligibility trace for ANCCR in **Fig S6A** (corresponds to γ=0.986), RPE at cue for the full state space will be 0.43 regardless of prior history.

However, there are still two major issues with TDRL. One is that TDRL RPE requires *a priori* knowledge of the true underlying state space to appropriately reflect the causal structure of the environment, while ANCCR works well without *a priori* knowledge of the environment. For standard state spaces assumed in neuroscience that reset their cue state representation on the next presentation of the cue, the RPE will be negative if a cue occurs between a cue-reward delay, a prediction that Test 8 violates. Further, the correct specification of state space becomes much harder for simple extensions of this task with multiple cue-outcome associations operating in parallel over different timescales (simulated in **FigS6B, C**). ANCCR works well under these conditions with no requirement of *a priori* specification of the state space.

Two is that the above calculation is the theoretical guarantee for RPE under correct specification of state space (discussed above) *and infinitely many timesteps of experience with a monotonically decreasing learning rate*, which ensures that all states in the state space are visited infinitely many times. In practice, even with the correct specification of state space, the online algorithm (as opposed to the theoretical one) is either exceedingly slow and fails to learn relationships (when learning rate monotonically decreases as 1/number of experienced time steps) or learns incorrect relationships with a constant learning rate (simulated above). We believe that this is, at least in part, due to the exponential growth in the state space size with respect to cue-reward delay (Number of states = $2^{X/dt} \approx 10^{18}$ for the parameters above in which ANCCR functions well in practice). One can in principle avoid this issue in TDRL by using functional approximation instead of tabular learning. However, learning the appropriate functional basis is itself hard, and is typically just assumed to be provided to the animal as an input. Hence, ANCCR is a parsimonious and sample-efficient explanation of animal learning that avoids the assumption that animals are *a priori* provided with the appropriate state space for value learning.

Table S1. (Statistical results):

Figure	Description	Test	Statistic	p value	Number of samples
Fig 2D	Probability of lick to CS	Paired t test	t = 4.42	Two-tailed p value = 3.0×10^{-4}	n = 20 iterations
Fig 2E	Probability of lick to CS2	Two sample t test	t = 7.32	Two-tailed p value = 9.2×10^{-7}	n = 20 iterations
Fig 2F	Probability of lick to CS1	Paired t test	t = 4.09	Two-tailed p value = 6.2×10^{-4}	n = 20 iterations
Fig 2H	Probability of Action 1	Paired t test	t = 7.06	Two-tailed p value = 1.0×10^{-6}	n = 20 iterations
Fig 2I	Probability of lick to CS2	Two sample t test	t = -6.33	Two-tailed p value = 2.0×10^{-7}	n = 20 iterations
Fig3C, right	Correlation between predicted DA and trial	One sample t test	t = RPE (CSC),-65.74; RPE (MS),-27.57; ANCCR, 18.60	Two-tailed p values = RPE (CSC), 1.7×10^{-83}; RPE (MS), 3.0×10^{-48}; ANCCR, 4.5×10^{-34}	n = 100 iterations
Fig 3E, right	Correlation between DA and trial	One sample t test	t = 4.40	Two-tailed p value = 0.0031	n = 8 animals
Fig3F, right	Correlation between predicted DA and previous IRI	One sample t test	t = RPE (CSC), -1.7×10^{3}; RPE (MS), -151.28; ANCCR, 335.03	Two-tailed p values = RPE (CSC), 5.0×10^{-223}; RPE (MS), 6.3×10^{-119}; ANCCR, 4.8×10^{-153}	n = 100 iterations
Fig 3G, right	Correlation between DA and previous IRI	One sample t test	t = 5.95	Two-tailed p value = 5.7×10^{-4}	n = 8 animals
Fig4F, I, L, left	Abruptness of change	Paired t test	t = Fig4F, 9.06; Fig4I, 22.92; Fig4L, 5.67	Two-tailed p values = Fig4F, 1.0×10^{-4}; Fig4I, 4.52×10^{-7}; Fig4L, 0.0013	n = 7 animals
Fig4F, L, right	Change trial	Paired t test	t = Fig4F, -2.93; Fig4L, -2.40	Two-tailed p values = Fig4F, 0.0263; Fig4L, 0.0531	n = 7 animals
Fig4I, right	ΔAIC	One sample t test	t = Lick, 7.49; DA, -0.86	Two-tailed p values = Lick, 2.9×10^{-4}; DA, 0.4244	n = 7 animals
Fig 4N	DA activity in background vs. extinction	Paired t test	t = -3.51	Two-tailed p value = 0.0126	n = 7 animals
Fig5B, left	Ratio of predicted cue response	One sample t test	t = RPE (CSC), -114.74; RPE (MS), -181.32; ANCCR, 322.53	Two-tailed p values = RPE (CSC), 4.1×10^{-107}; RPE (MS), 1.1×10^{-126}; ANCCR, 2.1×10^{-151}	n = 100 iterations
Fig5B, right	Ratio of predicted reward response	One sample t test	t = RPE (CSC), 87.67; RPE (MS), 62.86; ANCCR, 16.78	Two-tailed p values = RPE (CSC), 1.2×10^{-95}; RPE (MS), 1.3×10^{-81}; ANCCR, 1.1×10^{-30}	n = 100 iterations
Fig5D, left	Ratio of cue response	One sample t test	t = 6.64	Two-tailed p value = 5.6×10^{-4}	n = 7 animals
Fig5D, right	Ratio of reward response	One sample t test against null of 1	t = 2.95	Two-tailed p value = 0.0256	n = 7 animals
Fig6B	Difference of normalized DA response between early and late	Paired t test	t = -0.16	Two-tailed p value = 0.8771	n = 7 animals
Fig6D	Difference of normalized DA response between CS1 and CS2	Paired t test	t = -0.03	Two-tailed p value = 0.9752	n = 7 animals
Fig6G, H	Difference of normalized dopamine response during last 0.5s before reward (Fig6G) or 1s after reward (Fig6H) between DAT-Cre and WT groups	Two sample t test	t = left, -5.74 right, -3.99	Two-tailed p value = left, 1.3×10^{-4} right, 0.0021	n = 13 animals
Fig6H	Difference of CS1 response between DAT-	Two sample t test	t = RPE, -379.13	Two-tailed p values = RPE, 1.3×10^{-69}	n = 20 iterations

	Cre and WT groups		ANCCR, -10.22	ANCCR, 1.9×10^{-12}	
Fig6I	Normalized CS1 response (left) and behavior (right) across sessions in WT and DAT-Cre animals	Linear regression	t = CS1 response (left), WT, 5.49 DAT-Cre, 2.55 Behavior (right), WT, 3.61 DAT-Cre, 4.92	Two-tailed p values = CS1 response (left), WT, 4.35×10^{-6} DAT-Cre, 0.0166 Behavior (right), WT, 9.9×10^{-4} DAT-Cre, 3.40×10^{-5}	n = 35 sessions from 7 WT animals, 30 sessions from 6 DAT-Cre animals
Fig6I	Normalized CS1 response (left) and be-havior (right) at the last session between DAT-Cre and WT groups	Two sample t test	t = CS1 response (left), 2.80 Behavior (right), 0.88	Two-tailed p values = CS1 response (left), 0.0171 Behavior (right), 0.3999	n = 13 animals
FigS8B	Lick rate across first 7 sessions	Linear regression	t = Total, -0.79 Consum., -0.20 Non-consum., -0.58	Two-tailed p values = Total, 0.4389 Consum., 0.8397 Non-consum., 0.5618	n = 50 sessions from 8 animals
FigS8D	Correlation between DA and consummatory lick rate	One sample t test	t = 0.23	Two-tailed p value = 0.8245	n = 8 animals
FigS8E	Difference of DA activity between consummatory and non-consummatory bout	Paired t test	t = Consum vs. non-consum, 3.28;	Two-tailed p value = Consum vs. non-consum, 0.0134;	n = 8 animals
		One sample t test	t = Consum, 3.03; Non-consum, -2.55	Two-tailed p values = Consum, 0.0191; Non-consum, 0.0379	n = 8 animals
FigS8G	Correlation between baseline DA and session	One sample t test	t = 0.11	Two-tailed p value = 0.9190	n = 8 animals
FigS8K	Correlation between DA and IRI across first 7 sessions	Linear regression	t = -0.04	Two-tailed p value = 0.9707	n = 50 sessions from 8 animals
FigS8L	Lick rate across different IRI conditions	Linear regression	t = Consum, 1.00 Non-consum, -2.73	Two-tailed p values = Consum, 0.3338 Non-consum, 0.0147	n = 18 sessions from 6 animals
FigS10B	Difference of cumulative sum of DA reward response between Experiments 3 and 4	Paired t test	t = 3.37	Two-tailed p value = 0.0150	n = 7 animals
FigS11	Ratio of cue response	One sample t test	t = 10.42	Two-tailed p value = 4.6×10^{-5}	n = 7 animals
FigS12F	Anticipatory licks prior to CS2	Two-way mixed ANOVA	F = Genotype, 0.25 Session, 9.07 Genotype × session, 0.07	One-tailed p value = Genotype, 0.6241 Session, 0.0118 Genotype × session, 0.7915	n = 13 animals
FigS13C	Normalized predicted DA	Paired t test	t = T1 vs. Stand., 4.36; T2 vs. Stand., 23.42; T3 vs. Stand., 29.75	Two-tailed p values = T1 vs. Stand., 3.2×10^{-5}; T2 vs. Stand., 3.7×10^{-42}; T3 vs. Stand., 3.6×10^{-51}	n = 100 iterations
FigS13F	Normalized predicted DA	Paired t test	t = Slow vs. Stand., -129.07; Fast vs. Stand., 109.14	Two-tailed p values = Slow vs. Stand., 5.6×10^{-105}; Fast vs. Stand., 3.9×10^{-112}	n = 100 iterations

References and Notes

1. R. A. Rescorla, A. R. Wagner, “A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement” in *Classical conditioning II: Current Research and Theory*, pp. 64–99 (Appleton-Century-Crofts, 1972).

2. Y. Niv, G. Schoenbaum, Dialogues on prediction errors. *Trends Cogn. Sci.* **12**, 265–272 (2008). doi:10.1016/j.tics.2008.03.006 Medline

3. Y. Niv, Reinforcement learning in the brain. *J. Math. Psychol.* **53**, 139–154 (2009). doi:10.1016/j.jmp.2008.12.005

4. J. Y. Cohen, S. Haesler, L. Vong, B. B. Lowell, N. Uchida, Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* **482**, 85–88 (2012). doi:10.1038/nature10754 Medline

5. W. Schultz, P. Dayan, P. R. Montague, A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997). doi:10.1126/science.275.5306.1593 Medline

6. Y. K. Takahashi, M. R. Roesch, R. C. Wilson, K. Toreson, P. O’Donnell, Y. Niv, G. Schoenbaum, Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. *Nat. Neurosci.* **14**, 1590–1597 (2011). doi:10.1038/nn.2957 Medline

7. A. Mohebi, J. R. Pettibone, A. A. Hamid, J. T. Wong, L. T. Vinson, T. Patriarchi, L. Tian, R. T. Kennedy, J. D. Berke, Dissociable dopamine dynamics for learning and motivation. *Nature* **570**, 65–70 (2019). doi:10.1038/s41586-019-1235-y Medline

8. A. S. Hart, R. B. Rutledge, P. W. Glimcher, P. E. M. Phillips, Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. *J. Neurosci.* **34**, 698–704 (2014). doi:10.1523/JNEUROSCI.2489-13.2014 Medline

9. J. J. Day, M. F. Roitman, R. M. Wightman, R. M. Carelli, Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nat. Neurosci.* **10**, 1020–1028 (2007). doi:10.1038/nn1923 Medline

10. P. E. M. Phillips, G. D. Stuber, M. L. A. V. Heien, R. M. Wightman, R. M. Carelli, Subsecond dopamine release promotes cocaine seeking. *Nature* **422**, 614–618 (2003). doi:10.1038/nature01476 Medline

11. M. P. Saddoris, F. Cacciapaglia, R. M. Wightman, R. M. Carelli, Differential Dopamine Release Dynamics in the Nucleus Accumbens Core and Shell Reveal Complementary Signals for Error Prediction and Incentive Motivation. *J. Neurosci.* **35**, 11572–11582 (2015). doi:10.1523/JNEUROSCI.2344-15.2015 Medline

12. M. P. Saddoris, J. A. Sugam, G. D. Stuber, I. B. Witten, K. Deisseroth, R. M. Carelli, Mesolimbic dopamine dynamically tracks, and is causally linked to, discrete aspects of value-based decision making. *Biol. Psychiatry* **77**, 903–911 (2015). doi:10.1016/j.biopsych.2014.10.024 Medline

13. H. M. Bayer, P. W. Glimcher, Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129–141 (2005). doi:10.1016/j.neuron.2005.05.020 Medline

14. E. E. Steinberg, R. Keiflin, J. R. Boivin, I. B. Witten, K. Deisseroth, P. H. Janak, A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* **16**, 966–973 (2013). doi:10.1038/nn.3413 Medline

15. C. Y. Chang, G. R. Esber, Y. Marrero-Garcia, H.-J. Yau, A. Bonci, G. Schoenbaum, Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. *Nat. Neurosci.* **19**, 111–116 (2016). doi:10.1038/nn.4191 Medline

16. H.-C. Tsai, F. Zhang, A. Adamantidis, G. D. Stuber, A. Bonci, L. de Lecea, K. Deisseroth, Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science* **324**, 1080–1084 (2009). doi:10.1126/science.1168878 Medline

17. E. J. P. Maes, M. J. Sharpe, A. A. Usypchuk, M. Lozzi, C. Y. Chang, M. P. H. Gardner, G. Schoenbaum, M. D. Iordanova, Causal evidence supporting the proposal that dopamine transients function as temporal difference prediction errors. *Nat. Neurosci.* **23**, 176–178 (2020). doi:10.1038/s41593-019-0574-1 Medline

18. C. R. Gallistel, A. R. Craig, T. A. Shahan, Contingency, contiguity, and causality in conditioning: Applying information theory and Weber's Law to the assignment of credit problem. *Psychol. Rev.* **126**, 761–773 (2019). doi:10.1037/rev0000163 Medline

19. V. M. K. Namboodiri, J. M. Otis, K. van Heeswijk, E. S. Voets, R. A. Alghorazi, J. Rodriguez-Romaguera, S. Mihalas, G. D. Stuber, Single-cell activity tracking reveals that orbitofrontal neurons acquire and maintain a long-term memory to guide behavioral adaptation. *Nat. Neurosci.* **22**, 1110–1121 (2019). doi:10.1038/s41593-019-0408-1 Medline

20. V. M. K Namboodiri, G. D. Stuber, The learning of prospective and retrospective cognitive maps within neural circuits. *Neuron* **109**, 3552–3575 (2021). doi:10.1016/j.neuron.2021.09.034 Medline

21. V. M. K. Namboodiri, How do real animals account for the passage of time during associative learning? *Behav. Neurosci.* **136**, 383–391 (2022). doi:10.1037/bne0000516 Medline

22. J. R. Platt, Strong Inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science* **146**, 347–353 (1964). doi:10.1126/science.146.3642.347 Medline

23. G. Heymann, Y. S. Jo, K. L. Reichard, N. McFarland, C. Chavkin, R. D. Palmiter, M. E. Soden, L. S. Zweifel, Synergy of Distinct Dopamine Projection Populations in Behavioral Reinforcement. *Neuron* **105**, 909–920.e5 (2020). doi:10.1016/j.neuron.2019.11.024 Medline

24. W. Menegas, J. F. Bergan, S. K. Ogawa, Y. Isogai, K. Umadevi Venkataraju, P. Osten, N. Uchida, M. Watabe-Uchida, Dopamine neurons projecting to the posterior striatum form an anatomically distinct subclass. *eLife* **4**, e10032 (2015). doi:10.7554/eLife.10032 Medline

25. W. Menegas, K. Akiti, R. Amo, N. Uchida, M. Watabe-Uchida, Dopamine neurons projecting to the posterior striatum reinforce avoidance of threatening stimuli. *Nat. Neurosci.* **21**, 1421–1430 (2018). doi:10.1038/s41593-018-0222-1 Medline

26. S. Lammel, B. K. Lim, R. C. Malenka, Reward and aversion in a heterogeneous midbrain dopamine system. *Neuropharmacology* **76**, 351–359 (2014). doi:10.1016/j.neuropharm.2013.03.019 Medline

27. A. Lutas, H. Kucukdereli, O. Alturkistani, C. Carty, A. U. Sugden, K. Fernando, V. Diaz, V. Flores-Maldonado, M. L. Andermann, State-specific gating of salient cues by midbrain dopaminergic input to basal amygdala. *Nat. Neurosci.* **22**, 1820–1833 (2019). doi:10.1038/s41593-019-0506-0 Medline

28. B. T. Saunders, J. M. Richard, E. B. Margolis, P. H. Janak, Dopamine neurons create Pavlovian conditioned stimuli with circuit-defined motivational properties. *Nat. Neurosci.* **21**, 1072–1083 (2018). doi:10.1038/s41593-018-0191-4 Medline

29. A. A. Hamid, J. R. Pettibone, O. S. Mabrouk, V. L. Hetrick, R. Schmidt, C. M. Vander Weele, R. T. Kennedy, B. J. Aragona, J. D. Berke, Mesolimbic dopamine signals the value of work. *Nat. Neurosci.* **19**, 117–126 (2016). doi:10.1038/nn.4173 Medline

30. A. Lak, K. Nomoto, M. Keramati, M. Sakagami, A. Kepecs, Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision. *Curr. Biol.* **27**, 821–832 (2017). doi:10.1016/j.cub.2017.02.026 Medline

31. J. A. Parkinson, J. W. Dalley, R. N. Cardinal, A. Bamford, B. Fehnert, G. Lachenal, N. Rudarakanchana, K. M. Halkerston, T. W. Robbins, B. J. Everitt, Nucleus accumbens dopamine depletion impairs both acquisition and performance of appetitive Pavlovian approach behaviour: implications for mesoaccumbens dopamine function. *Behavioural Brain Research*. **137**, 149–163 (2002). doi:10.1523/JNEUROSCI.2344-15.2015 Medline

32. M. Darvas, A. M. Wunsch, J. T. Gibbs, R. D. Palmiter, Dopamine dependency for acquisition and performance of Pavlovian conditioned response. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2764–2769 (2014). doi:10.1073/pnas.1400332111 Medline

33. K. Yamaguchi, Y. Maeda, T. Sawada, Y. Iino, M. Tajiri, R. Nakazato, S. Ishii, H. Kasai, S. Yagishita, A behavioural correlate of the synaptic eligibility trace in the nucleus accumbens. *Sci. Rep.* **12**, 1921 (2022). doi:10.1038/s41598-022-05637-6 Medline

34. M. G. Kutlu, J. E. Zachry, P. R. Melugin, S. A. Cajigas, M. F. Chevee, S. J. Kelly, B. Kutlu, L. Tian, C. A. Siciliano, E. S. Calipari, Dopamine release in the nucleus accumbens core signals perceived saliency. *Curr. Biol.* **31**, 4748–4761.e8 (2021). doi:10.1016/j.cub.2021.08.052 Medline

35. T. Patriarchi, J. R. Cho, K. Merten, M. W. Howe, A. Marley, W.-H. Xiong, R. W. Folk, G. J. Broussard, R. Liang, M. J. Jang, H. Zhong, D. Dombeck, M. von Zastrow, A. Nimmerjahn, V. Gradinaru, J. T. Williams, L. Tian, Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors. *Science* **360**, eaat4422 (2018). doi:10.1126/science.aat4422 Medline

36. N. D. Daw, A. C. Courville, D. S. Touretzky, D. S. Touretzky, Representation and timing in theories of the dopamine system. *Neural Comput.* **18**, 1637–1677 (2006). doi:10.1162/neco.2006.18.7.1637 Medline

37. H.-E. Tan, A. C. Sisti, H. Jin, M. Vignovich, M. Villavicencio, K. S. Tsang, Y. Goffer, C. S. Zuker, The gut-brain axis mediates sugar preference. *Nature* **580**, 511–516 (2020). doi:10.1038/s41586-020-2199-7 Medline

38. C. R. Gallistel, T. A. Mark, A. P. King, P. E. Latham, The rat approximates an ideal detector of changes in rates of reward: Implications for the law of effect. *J. Exp. Psychol. Anim. Behav. Process.* **27**, 354–372 (2001). doi:10.1037/0097-7403.27.4.354 Medline

39. R. A. Rescorla, Pavlovian conditioning and its proper control procedures. *Psychol. Rev.* **74**, 71–80 (1967). doi:10.1037/h0024109 Medline

40. R. A. Rescorla, Probability of shock in the presence and absence of CS in fear conditioning. *J. Comp. Physiol. Psychol.* **66**, 1–5 (1968). doi:10.1037/h0025984 Medline

41. C. R. Gallistel, Robert Rescorla: Time, Information and Contingency. *Rev. Hist. Psicol.* **42**, 7–21 (2021).

42. V. M. K Namboodiri, T. Hobbs, I. Trujillo-Pisanty, R. C. Simon, M. M. Gray, G. D. Stuber, Relative salience signaling within a thalamo-orbitofrontal circuit governs learning rate. *Curr. Biol.* **31**, 5176–5191.e5 (2021). doi:10.1016/j.cub.2021.09.037 Medline

43. C. D. Fiorillo, W. T. Newsome, W. Schultz, The temporal precision of reward prediction in dopamine neurons. *Nat. Neurosci.* **11**, 966–973 (2008). doi:10.1038/nn.2159 Medline

44. A. Pastor-Bernier, A. Stasiak, W. Schultz, Reward-specific satiety affects subjective value signals in orbitofrontal cortex during multicomponent economic choice. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2022650118 (2021). doi:10.1073/pnas.2022650118 Medline

45. J. R. Hollerman, W. Schultz, Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.* **1**, 304–309 (1998). doi:10.1038/1124 Medline

46. S. Kobayashi, W. Schultz, Influence of reward delays on responses of dopamine neurons. *J. Neurosci.* **28**, 7837–7846 (2008). doi:10.1523/JNEUROSCI.1600-08.2008 Medline

47. M. E. Bouton, S. Maren, G. P. McNally, Behavioral and neurobiological mechanisms of pavlovian and instrumental extinction learning. *Physiol. Rev.* **101**, 611–681 (2021). doi:10.1152/physrev.00016.2020 Medline

48. W.-X. Pan, R. Schmidt, J. R. Wickens, B. I. Hyland, Tripartite mechanism of extinction suggested by dopamine neuron activity and temporal difference model. *J. Neurosci.* **28**, 9619–9631 (2008). doi:10.1523/JNEUROSCI.0255-08.2008 Medline

49. W. Zhong, Y. Li, Q. Feng, M. Luo, Learning and Stress Shape the Reward Response Patterns of Serotonin Neurons. *J. Neurosci.* **37**, 8863–8875 (2017). doi:10.1523/JNEUROSCI.1181-17.2017 Medline

50. R. Amo, S. Matias, A. Yamanaka, K. F. Tanaka, N. Uchida, M. Watabe-Uchida, A gradual temporal shift of dopamine responses mirrors the progression of temporal difference error in machine learning. *Nat. Neurosci.* **25**, 1082–1092 (2022). doi:10.1038/s41593-022-01109-2 Medline

51. K. E. Bouchard, M. S. Brainard, Neural encoding and integration of learned probabilistic sequences in avian sensory-motor circuitry. *J. Neurosci.* **33**, 17710–17723 (2013). doi:10.1523/JNEUROSCI.2181-13.2013 Medline

52. Y. Komura, R. Tamura, T. Uwano, H. Nishijo, K. Kaga, T. Ono, Retrospective and prospective coding for predicted reward in the sensory thalamus. *Nature* **412**, 546–549 (2001). doi:10.1038/35087595 Medline

53. H. E. Manzur, K. Vlasov, S.-C. Lin, A retrospective and stepwise learning strategy revealed by neuronal activity in the basal forebrain. bioRxiv 2022.04.01.486795 [Preprint] (2022); doi:10.1101/2022.04.01.486795v1.

54. L. T. Coddington, J. T. Dudman, The timing of action determines reward prediction signals in identified midbrain dopamine neurons. *Nat. Neurosci.* **21**, 1563–1573 (2018). doi:10.1038/s41593-018-0245-7 Medline

55. K. Lee, L. D. Claar, A. Hachisuka, K. I. Bakhurin, J. Nguyen, J. M. Trott, J. L. Gill, S. C. Masmanidis, Temporally restricted dopaminergic control of reward-conditioned movements. *Nat. Neurosci.* **23**, 209–216 (2020). doi:10.1038/s41593-019-0567-0 Medline

56. B. Engelhard, J. Finkelstein, J. Cox, W. Fleming, H. J. Jang, S. Ornelas, S. A. Koay, S. Y. Thiberge, N. D. Daw, D. W. Tank, I. B. Witten, Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. *Nature* **570**, 509–513 (2019). doi:10.1038/s41586-019-1261-9 Medline

57. R. N. Hughes, K. I. Bakhurin, E. A. Petter, G. D. R. Watson, N. Kim, A. D. Friedman, H. H. Yin, Ventral Tegmental Dopamine Neurons Control the Impulse Vector during Motivated Behavior. *Curr. Biol.* **30**, 2681–2694.e5 (2020). doi:10.1016/j.cub.2020.05.003 Medline

58. H. R. Kim, A. N. Malik, J. G. Mikhael, P. Bech, I. Tsutsui-Kimura, F. Sun, Y. Zhang, Y. Li, M. Watabe-Uchida, S. J. Gershman, N. Uchida, A Unified Framework for Dopamine Signals across Timescales. *Cell* **183**, 1600–1616.e25 (2020). doi:10.1016/j.cell.2020.11.013 Medline

59. A. A. Hamid, M. J. Frank, C. I. Moore, Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment. *Cell* **184**, 2733–2749.e16 (2021). doi:10.1016/j.cell.2021.03.046 Medline

60. M. J. Sharpe, C. Y. Chang, M. A. Liu, H. M. Batchelor, L. E. Mueller, J. L. Jones, Y. Niv, G. Schoenbaum, Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nat. Neurosci.* **20**, 735–742 (2017). doi:10.1038/nn.4538 Medline

61. M. J. Sharpe, H. M. Batchelor, L. E. Mueller, C. Yun Chang, E. J. P. Maes, Y. Niv, G. Schoenbaum, Dopamine transients do not act as model-free prediction errors during associative learning. *Nat. Commun.* **11**, 106 (2020). doi:10.1038/s41467-019-13953-1 Medline

62. B. M. Seitz, I. B. Hoang, L. E. DiFazio, A. P. Blaisdell, M. J. Sharpe, Dopamine errors drive excitatory and inhibitory components of backward conditioning in an outcome-specific manner. *Curr. Biol.* **32**, 3210–3218.e3 (2022). doi:10.1016/j.cub.2022.06.035 Medline

63. I. Trujillo-Pisanty, K. Conover, P. Solis, D. Palacios, P. Shizgal, Dopamine neurons do not constitute an obligatory stage in the final common path for the evaluation and pursuit of brain stimulation reward. *PLOS ONE* **15**, e0226722 (2020). doi:10.1371/journal.pone.0226722 Medline

64. W. Z. Goh, V. Ursekar, M. W. Howard, Predicting the Future With a Scale-Invariant Temporal Memory for the Past. *Neural Comput.* **34**, 642–685 (2022). doi:10.1162/neco_a_01475 Medline

65. K. H. Shankar, M. W. Howard, A scale-invariant internal representation of time. *Neural Comput.* **24**, 134–193 (2012). doi:10.1162/NECO_a_00212 Medline

66. A. Tsao, J. Sugar, L. Lu, C. Wang, J. J. Knierim, M.-B. Moser, E. I. Moser, Integrating time from experience in the lateral entorhinal cortex. *Nature* **561**, 57–62 (2018). doi:10.1038/s41586-018-0459-6 Medline

67. W. Wei, A. Mohebi, J. D. Berke, Striatal dopamine pulses follow a temporal discounting spectrum. bioRxiv 2021.10.31.466705 [Preprint] (2021); doi:10.1101/2021.10.31.466705.

68. T. J. Madarasz, L. Diaz-Mataix, O. Akhand, E. A. Ycu, J. E. LeDoux, J. P. Johansen, Evaluation of ambiguous associations in the amygdala by learning the structure of the environment. *Nat. Neurosci.* **19**, 965–972 (2016). doi:10.1038/nn.4308 Medline

69. S. J. Gershman, Y. Niv, Exploring a latent cause theory of classical conditioning. *Learn. Behav.* **40**, 255–268 (2012). doi:10.3758/s13420-012-0080-8 Medline

70. P. D. Balsam, C. R. Gallistel, Temporal maps and informativeness in associative learning. *Trends Neurosci.* **32**, 73–78 (2009). doi:10.1016/j.tins.2008.10.004 Medline

71. S. J. Gershman, D. M. Blei, Y. Niv, Context, learning, and extinction. *Psychol. Rev.* **117**, 197–209 (2010). doi:10.1037/a0017808 Medline

72. E. C. J. Syed, L. L. Grima, P. J. Magill, R. Bogacz, P. Brown, M. E. Walton, Action initiation shapes mesolimbic dopamine encoding of future rewards. *Nat. Neurosci.* **19**, 34–36 (2016). doi:10.1038/nn.4187 Medline

73. A. L. Collins, V. Y. Greenfield, J. K. Bye, K. E. Linker, A. S. Wang, K. M. Wassum, Dynamic mesolimbic dopamine signaling during action sequence learning and expectation violation. *Sci. Rep.* **6**, 20231 (2016). doi:10.1038/srep20231 Medline

74. A. Guru, C. Seo, R. J. Post, D. S. Kullakanda, J. A. Schaffer, M. R. Warden, Ramping activity in midbrain dopamine neurons signifies the use of a cognitive map. bioRxiv2020.05.21.108886 [Preprint] (2020), doi:10.1101/2020.05.21.108886.

75. N. G. Hollon, E. W. Williams, C. D. Howard, H. Li, T. I. Traut, X. Jin, Nigrostriatal dopamine signals sequence-specific action-outcome prediction errors. *Curr. Biol.* **31**, 5350–5363.e5 (2021). doi:10.1016/j.cub.2021.09.040 Medline

76. W. van Elzelingen, P. Warnaar, J. Matos, W. Bastet, R. Jonkman, D. Smulders, J. Goedhoop, D. Denys, T. Arbab, I. Willuhn, Striatal dopamine signals are region specific and temporally stable across action-sequence habit formation. *Curr. Biol.* **32**, 1163–1174.e6 (2022). doi:10.1016/j.cub.2021.12.027 Medline

77. J. P. Gavornik, M. G. H. Shuler, Y. Loewenstein, M. F. Bear, H. Z. Shouval, Learning reward timing in cortex through reward dependent expression of synaptic plasticity. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 6826–6831 (2009). doi:10.1073/pnas.0901835106 Medline

78. V. M. K. Namboodiri, S. Mihalas, T. M. Marton, M. G. Hussain Shuler, A general theory of intertemporal decision-making and the perception of time. *Front. Behav. Neurosci.* **8**, 61 (2014). doi:10.3389/fnbeh.2014.00061 Medline

79. P. N. Tobler, C. D. Fiorillo, W. Schultz, Adaptive coding of reward value by dopamine neurons. *Science* **307**, 1642–1645 (2005). doi:10.1126/science.1105370 Medline

80. N. F. Parker, C. M. Cameron, J. P. Taliaferro, J. Lee, J. Y. Choi, T. J. Davidson, N. D. Daw, I. B. Witten, Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. *Nat. Neurosci.* **19**, 845–854 (2016). doi:10.1038/nn.4287 Medline

81. H. Jeong, A. Taylor, J. R. Floeder, M. Lohmann, S. Mihalas, B. Wu, M. Zhou, D. A. Burke, V. M. K Namboodiri, Mesolimbic dopamine release conveys causal associations, Version 1, Zenodo (2022); https://zenodo.org/record/7302777#.Y4j503bMI2w

82. J. Pearl, Causal Diagrams for Empirical Research. *Biometrika* **82**, 669–688 (1995). doi:10.1093/biomet/82.4.669

83. C. R. Gallistel, J. Gibbon, Time, rate, and conditioning. *Psychol. Rev.* **107**, 289–344 (2000). doi:10.1037/0033-295X.107.2.289 Medline

84. I. M. Bright, M. L. R. Meister, N. A. Cruzado, Z. Tiganj, E. A. Buffalo, M. W. Howard, A temporal record of the past with a spectrum of time constants in the monkey entorhinal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 20274–20283 (2020). doi:10.1073/pnas.1917197117 Medline

85. Y. M. Ulrich-Lai, A. M. Christiansen, M. M. Ostrander, A. A. Jones, K. R. Jones, D. C. Choi, E. G. Krause, N. K. Evanson, A. R. Furay, J. F. Davis, M. B. Solomon, A. D. de Kloet, K. L. Tamashiro, R. R. Sakai, R. J. Seeley, S. C. Woods, J. P. Herman, Pleasurable behaviors reduce stress via brain reward pathways. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 20529–20534 (2010). doi:10.1073/pnas.1007740107 Medline

86. E. N. Holly, K. A. Miczek, Ventral tegmental area dopamine revisited: Effects of acute and repeated stress. *Psychopharmacology* **233**, 163–186 (2016). doi:10.1007/s00213-015-4151-3 Medline

87. C. E. Stelly, S. C. Tritley, Y. Rafati, M. J. Wanat, Acute Stress Enhances Associative Learning via Dopamine Signaling in the Ventral Lateral Striatum. *J. Neurosci.* **40**, 4391–4400 (2020). doi:10.1523/JNEUROSCI.3003-19.2020 Medline

88. D. George, R. V. Rikhye, N. Gothoskar, J. S. Guntupalli, A. Dedieu, M. Lázaro-Gredilla, Clone-structured graph representations enable flexible learning and vicarious evaluation of cognitive maps. *Nat. Commun.* **12**, 2392 (2021). doi:10.1038/s41467-021-22559-5 Medline

89. J. C. R. Whittington, T. H. Muller, S. Mark, G. Chen, C. Barry, N. Burgess, T. E. J. Behrens, The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell* **183**, 1249–1263.e23 (2020). doi:10.1016/j.cell.2020.10.024 Medline

90. B. Hayden, Y. Niv, The case against economic values in the brain. PsyArXiv 10.31234/osf.io/7hgup [Preprint] (2020), doi:10.31234/osf.io/7hgup.

91. F. Etscorn, R. Stephens, Establishment of conditioned taste aversions with a 24-hour CS-US interval. *Physiol. Psychol.* **1**, 251–253 (1973). doi:10.3758/BF03326916

92. T. Akam, M. E. Walton, pyPhotometry: Open source Python based hardware and software for fiber photometry data acquisition. *Sci. Rep.* **9**, 3521 (2019). doi:10.1038/s41598-019-39724-y Medline

93. T. N. Lerner, C. Shilyansky, T. J. Davidson, K. E. Evans, K. T. Beier, K. A. Zalocusky, A. K. Crow, R. C. Malenka, L. Luo, R. Tomer, K. Deisseroth, Intact-Brain Analyses Reveal Distinct Information Carried by SNc Dopamine Subcircuits. *Cell* **162**, 635–647 (2015). doi:10.1016/j.cell.2015.07.014 Medline

94. E. Martianova, S. Aronson, C. D. Proulx, Fiber Photometry to Record Neural Activity in Freely-Moving Animals, *J. Vis. Exp.* **152**, e60278 (2019). doi:10.3791/60278 Medline

95. E. A. Ludvig, R. S. Sutton, E. J. Kehoe, Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Comput.* **20**, 3034–3054 (2008). doi:10.1162/neco.2008.11-07-654 Medline

96. J. Pearl, Causal inference in statistics: An overview. *Stat. Surv.* **3**, 96–146 (2009). doi:10.1214/09-SS057