Lateral habenula neurons signal errors in the prediction of reward information
Ethan S. Bromberg-Martin and Okihide Hikosaka

Supplementary Figures

1. Computational TD learning model of IPEs and conventional RPEs
2. Test of computational mechanisms for assigning value to informative cues
3. Behavioral performance and neural activity for each recording session
4. No significant difference in mean response to forced vs. choice trials
5. Trend for different response time course on forced vs. choice trials
6. Conventional RPE signals in the information-predictive and remaining neurons
7. Absence of consistent IPE signals in the remaining neurons
8. A further test of IPE and cRPE transmission by single neurons
9. Neural data are consistent with low-info choices being caused by noisy action selection
10. Signals related to IPEs and cRPEs in four example neurons – summary
11. Signals related to IPEs and cRPEs in four example neurons – raw activity

Full description for Supplementary Fig. 1:
Computational TD learning model of IPEs and conventional RPEs

Here we show that the information-related effects seen in the neural and behavioral data (**Fig. S1a**) cannot be accounted for by a conventional temporal difference (TD) learning model of RPEs[51] (**Fig. S1c**) but can be accounted for by a modified model that assigns bonus reward value to the act of viewing informative cues (**Fig. S1b**). In the main text, we use the simulated neural activity from these TD models to make the diagrams showing theoretical IPE signals (**Fig. 4-6**). We next describe the formal representation of the task, the TD learning models, the procedure for comparing the models to the data, and the results of the comparison.

Formal representation of the task:

In order to apply TD models to the data we had to represent the task as a Markov decision process (MDP): a sequence of discrete states that transition to each other with specified probabilities (potentially dependent on the subject's actions) and that deliver specified reward outcomes[52]. We represented each trial with a series of state transitions representing the task epochs: fixation → target array onset → target chosen → visual cue → end of trial. Water rewards were delivered on the transition from the visual cue to the end of the trial. Each epoch had several potential states matching the potential stimuli during the task: five target arrays (choice 100% vs. 50%, choice 50% vs. 0%, forced 100%, forced 50%, forced 0%), three chosen targets (100%, 50%, and 0%), and four visual cues (info-big, info-small, random 1, random 2). The transition probabilities between states (and their associated rewards) were set to match the true task structure. For example, the target arrays transitioned to the possible chosen targets based on the model's behavioral preferences (as described below), and the probability of each target array being presented was set equal to its true probability of being presented if the subject had had those behavioral preferences (Methods). The reward sizes were set equal to the true amount of water reward in milliliters (Methods).

Temporal difference models:

We used a simple TD model to estimate the reward value of each event during our behavioral task, to generate an RPE signal, and make choices[51, 52]. In TD learning, the value V(s) of a state s is equal to the expected future amount of time-discounted primary reward. This can be calculated by considering all possible state transitions from state s and taking into account the amount of immediate primary reward r and the value of the next state s'. The relative value of immediate vs. delayed rewards is controlled by a temporal discounting parameter γ. Since our task did not manipulate reward timing this parameter did not influence the results, so for simplicity we fixed it at $\gamma = 1$ (other settings produced similar results). To formalize this, we represent each potential state transition using a tuple (s, s', r, p), meaning that state s transitions to state s' while delivering reward r with probability p. Let $T(s)$ be the set of all state transitions from state s. Then the value of state s is defined by the equation:

$$V(s) = \Sigma_{(s,\, s',\, r,\, p) \,\in\, T(s)}\, p(\gamma V(s') + r).$$

The RPE signal in TD learning (called the TD error, δ), is triggered by each state transition and is equal to the difference between the predicted reward value of the current state, V(s), and the reward value delivered by the transition, $\gamma V(s') + r$. It is controlled by the equation:

$$\delta = (\gamma V(s') + r) - V(s)$$

We calculated the values of all of the states by initializing V(end of trial) = 0 and then recursively calculating the remaining values (i.e., using an 'episodic' TD model[52]). The model chose between actions using the standard softmax rule, so that the sensitivity of the subject's choice to the action values was controlled by the parameter β. Specifically, if the values of the targets 1 and 2 when presented alone on forced trials were V_1 and V_2, then the probability of choosing target 1 over target 2 on a choice trial was set equal to:

$$\Pr(\text{choose target 1}) = \exp(\beta V_1) / (\exp(\beta V_1) + \exp(\beta V_2)).$$

In order to compare the model with the neural data, the model was given an additional scaling parameter *k* that controlled the mapping from model RPEs to neural activity. That is, the model's predicted neural RPE effects (in units of spikes/s) were set equal to *k* times the model's RPE signal (in units of milliliters of water). Note that habenula neurons signal RPEs in an inverted manner so *k* should be negative.

In summary, the *conventional TD model* had two parameters:

1. The softmax parameter β, controlling choices.
2. The scaling parameter *k*, mapping from RPEs to neural activity.

We also implemented a modified TD model, called the *information-bonus TD model*, to represent our hypothesis that the brain assigns additional reward value to the act of viewing informative cues. This was identical to the conventional model except that it received a bonus reward r_{info} upon each transition to an informative cue state. (Note that similar results can be produced by giving a bonus to only one of the two informative cues, or by giving a penalty to the random cues). So this model had a third parameter:

3. The bonus parameter r_{info}, setting the bonus reward for informative cues.

Procedure for comparing the models to the data:

For each model, we found the parameters that optimized the fit between the model and the neural and behavioral data. The data were represented by a vector **x** of 11 values representing 9 neural IPE and cRPE effects (measured as firing rate differences between pairs of task conditions) and 2 behavioral effects (measured as choice percentages), as follows:

Effect name	*Effect measurement*	*Fig. #*
Negative IPE, target	(0% info) – (50% info)	4a
Negative IPE, cue	(unpredicted no-info) – (predicted no-info)	5a
Positive IPE, target	(100% info) – (50% info)	4a
Positive IPE, small cue	(unpredicted info, small) – (predictable info, small)	6a
Positive IPE, big cue	(unpredicted info, big) – (predictable info, big)	6a
Negative cRPE, cue	(predicted info-small cue) – (predicted random cues)	2a
Negative cRPE, reward	(random-small reward) – (info-small reward)	2a
Positive cRPE, cue	(predicted info-big cue) – (predicted random cues)	2a

Positive cRPE, reward	(random-big reward) – (info-big reward)	2a
Choice type #1	percentage of choosing 100% > 50% info	1c
Choice type #2	percentage choosing 50% > 0% info	1c

The fit was optimized by using the matlab function 'fminsearch' to minimize the fitting error. The fitting error was defined as the sum of the squared errors between the measured mean effects in the data and the predicted effects from the model, normalized by the variability of the measurements (defined for each measurement as the square of the standard error of the mean). Thus, if the i-th effect had a measured mean of $x(i)$ with a standard error of $\sigma(i)$, and a model-predicted effect of $y(i)$, then the fitting error was given by the equation: $\Sigma_i \, (y(i) - x(i))^2 / \sigma(i)^2$. This ensured that the fit was based on all 11 effects while giving greater weight to the effects that could be measured more reliably in the data.

Generation of hypothetical response diagrams in main text:

To generate the hypothetical response diagrams in the main text, we generated neural responses to each task event based on the fitted firing rate response for that event. Responses were drawn as triangular waveforms such that the mean of the triangle was equal to the mean neural response (i.e., the peak of the triangle was 2x as high as the mean neural response).

Results:

True data (**Fig. S1a**)

As expected from our results in the main text, the population of information-predictive neurons was significantly excited for all negative IPEs and significantly inhibited for all positive IPEs (top panel, red bars show IPE effects in cross-validated data; signed-rank test, all $P < 0.05$). The population was also excited for all negative cRPEs and inhibited for all positive cRPEs (middle panel, black bars; signed-rank test, all $P < 0.05$). The IPE signals were smaller than the cRPE signals but had a similar overall pattern. Thus, the IPE signals had a close resemblance to the cRPE signals if they were scaled by a factor of 5 (middle panel, red bars), suggesting that the neurons might have assigned reward value to the informative cues equal to approximately $1/5^{th}$ the value difference between the big and small water rewards.

Conventional TD model (**Fig. S1c**), fitted parameters: $\beta = 20$, $k = -30$.

The conventional TD model produced cRPE signals similar to those observed in the data, with excitation for negative cRPEs and inhibition for positive cRPEs (middle panel, black bars). However, the model could not reproduce the observed pattern of IPE signals – it had no differential response during information-related events (top and middle panels, red bars), and had no behavioral preference between the targets (bottom panel).

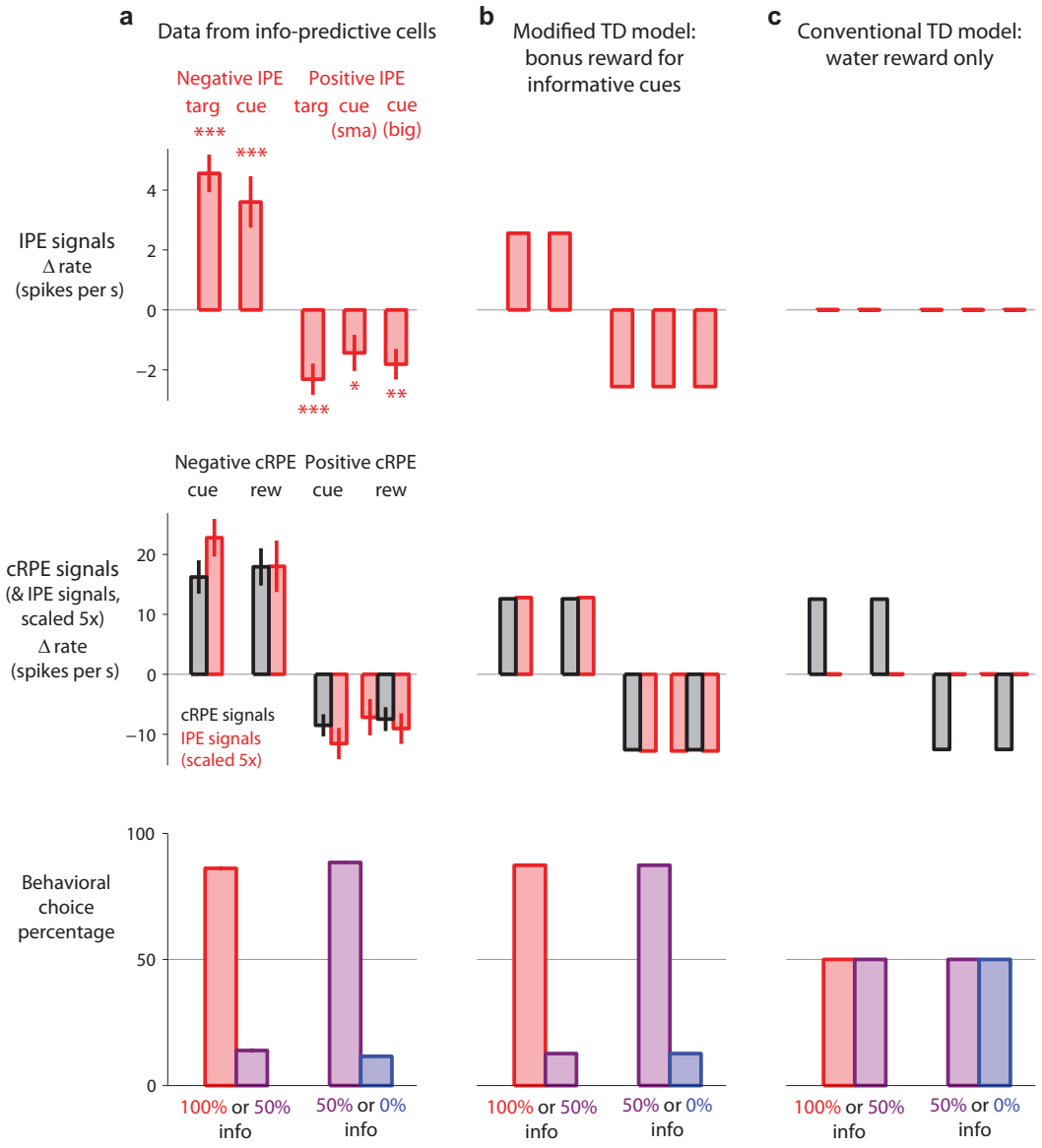
All five IPE signals in the data and both behavioral preference measurements were significantly different from those for the model (all $P < 0.05$, signed-rank test for IPE effects and binomial test for choice percentages).

Information-bonus model (**Fig. S1b**), fitted parameters: $\beta = 23$, $k = -30$, $r_{info} = 0.17$.

The information-bonus model produced both IPE and cRPE signals similar to those observed in the data. It had IPE signals with the correct direction and similar magnitudes to those seen in the data (top panel). It had behavioral choices closely matching those seen in the data

(bottom panel). And it had cRPE signals resembling those seen in the data, and a similar proportional relationship between IPEs and cRPEs (middle panel). Notably, the model assigned reward value to the informative cues equal to exactly $1/5^{th}$ the value difference between the big and small water rewards (r_{info} /(0.88-0.04) = 0.20), matching the pattern seen in the neural data.

Most of the IPE signals in the data and behavioral preference measurements were not significantly different from the model ($P > 0.05$, signed-rank test for IPE effects and binomial test for choice percentages), although one signal was slightly stronger than the model (negative IPE for the targets, $P = 0.006$, signed-rank test) and one was slightly weaker than the model (positive IPE for unpredicted info with the big reward cue, $P = 0.035$, signed-rank test).



Supplementary Figure 1. Computational TD learning model of IPEs and conventional RPEs. The true neural and behavioral effects in the cross-validated data, used to fit the models (**a**) are similar to a modified TD model that receives bonus reward for viewing informative cues (**b**), but cannot be accounted for by conventional TD models that only receive water rewards (**c**).
Top row: Red bars: neural IPE signals measured as the change in firing rate induced by negative IPEs (0% info target, unpredicted no-info) and positive IPEs (100% info target, unpredicted info during small reward trials and during big reward trials).
Middle row: Black bars: neural cRPE signals measured as the change in firing rate induced by negative cRPEs (Info-small cue, unpredicted small reward) and positive cRPEs (Info-big cue, unpredicted big rewards). Red bars: the same IPE signals as in the top row but scaled up by a factor of 5, revealing that the cRPE and IPE signals have a comparable response pattern despite their different magnitudes.
Bottom row: Behavioral choice percentages between the information-related targets.

Full description for Supplementary Fig. 2:
Test of computational mechanisms for assigning value to informative cues

Having shown that our data can be accounted for by a model that assigns bonus value to informative cues relative to random cues, we next asked whether our data can constrain the possible mechanisms by which this bonus value is assigned.

In this section we show that our data provides evidence against one potential mechanism for assigning greater value to the informative cues, called a disengagement mechanism, which was proposed in a recent TD learning model[43]. The disengagement model can reproduce the correct pattern of IPE and cRPE signals seen in the average neural activity (**Fig. S2a**). However, it also makes a prediction about the trial-to-trial reliability of neural responses that is not borne out in the data. The model predicts that animals frequently 'disengage' from the task and forget their reward predictions. This would put sharp limits on the ability of neurons to discriminate between predictable and unpredictable reward outcomes. In the data, however, most habenula neurons had very strong discrimination between predictable and unpredictable rewards, beyond the limits predicted by the model (**Fig. S2b**).

Disengagement model:

The disengagement model is the same as conventional TD learning but is modified to have an internal state of 'disengagement', representing a time period when the subject loses track of the current task state (e.g. by not paying attention to the task) and is unable to predict rewards[43]. The model engages at the start of each trial and then has a chance of transitioning to the disengaged state at each moment during the task. The disengaged state has its value fixed at zero, so a transition to the disengaged state causes a negative RPE that punishes previous states and actions. The chance of disengagement depends on the value of the current state – it occurs often during low-value states and rarely during high-value states.

The idea of this mechanism is to attenuate the values of all states but to have a weaker effect on informative cue states than random cue states, thus leaving the informative cue states with a relatively higher value. Specifically, the "info-big cue" state has a high value so disengagement rarely occurs. The "info-small cue" state has a low value so disengagement occurs often, but its value is already near zero so it cannot fall very far. But the "random cue" state has an intermediate value, low enough that disengagement occurs fairly often but high enough that its value has far to fall.

To test whether this model could account for our data, we implemented it as follows. The probability of disengagement per second of time spent in state s with value $V(s)$ is denoted as $\varepsilon(s)$ and is controlled by the equation[43]:

$$\varepsilon(s) = \varepsilon_0 \exp(-V(s)\psi)$$

Here the parameter ε_0 controls the rate of disengagement and the parameter ψ controls how the rate of disengagement is influenced by the current state's value. Given the state's disengagement rate, the probability that the state completes successfully without disengaging (and therefore is able to update its value in the normal way based on future states and rewards) is controlled by the equation[43]:

$$\Pr(\text{normal value update} \mid s) = (1 - \varepsilon(s))^{\tau}$$

Where τ is the duration of the state in seconds. The final result of these disengagement effects is that the value of a state is based on the fraction of trials when its value can be updated normally (since the remaining trials lead to disengagement which always has a value of zero). This can be expressed with the equation:

$$V(s) = V_{conv}(s)\Pr(\text{normal value update} \mid s)$$

Where $V_{conv}(s)$ is the value that would be given by the conventional updating equation, i.e. $V_{conv}(s) = \Sigma_{(s, s', r, p) \in T(s)}\, p(\gamma V(s') + r)$. Note that by expanding terms we can see that $V(s)$ in this model is defined recursively: $V(s) = V_{conv}(s)(1 - \varepsilon_0 \exp(-V(s)\psi))^{\tau}$. We therefore solve for the state value numerically by using an iterative procedure to find the unique setting of $V(s)$ that satisfies the equation.

In summary, the disengagement model had four parameters:

1. The softmax parameter β, controlling choices.
2. The scaling parameter k, mapping from RPEs to neural activity.
3. The parameter ε_0 setting the overall rate of disengagement.
4. The parameter ψ setting the sensitivity of disengagement to state values.

Model bounds on ROC area for neural discrimination:

The disengagement model also makes a prediction about the trial-to-trial reliability of neural responses[43]. According to the model, neural responses during the disengaged state reflect only the immediate delivery of reward without being influenced by prior predictions[43] (because the disengaged state has a predicted reward value of zero). So the model's disengagement rate puts a strict limit on its ability to discriminate between predictable and unpredictable reward delivery. In our task, this corresponds to a strict bound on the ability of neurons to discriminate whether a reward was delivered on an informative cue trial (predictable) or a random cue trial (unpredictable).

The precise bounds are set by the probability that the model is in the disengaged state at the moment when reward is delivered. We calculated this separately for info-big, info-small, and random cue trials (here denoted as $p_d(IB)$,$p_d(IS)$, and $p_d(R)$). As it turned out, the best-fitting model disengaged very often – its probabilities were $p_d(IB) = 0.27$, $p_d(R) = 0.55$, $p_d(IS) = 0.91$. This is roughly comparable to (but slightly higher than) the disengagement probabilities for the originally proposed parameter settings of the model[43].

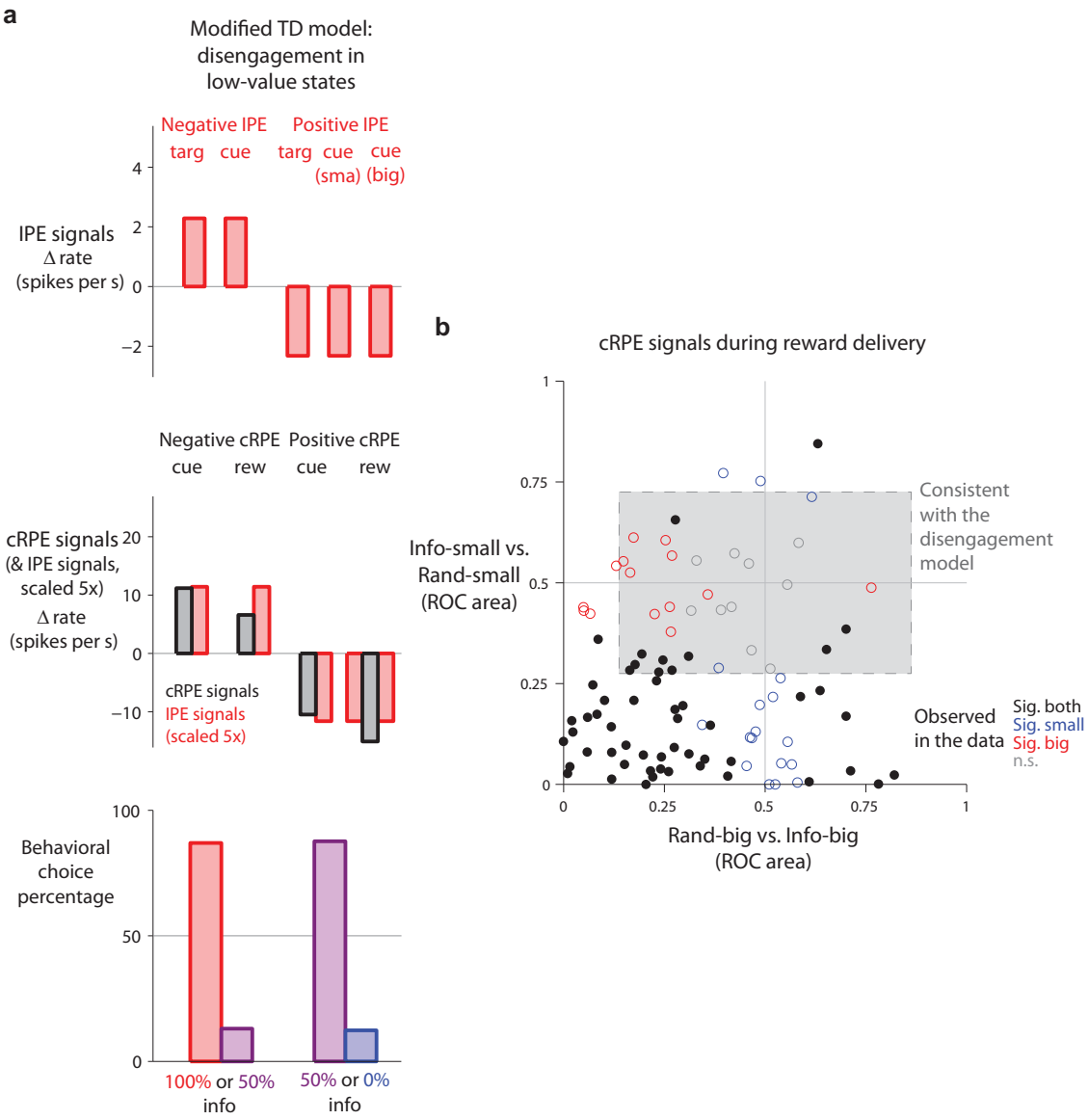
The bounds on neural discrimination were then calculated as follows. We first consider the ROC area for discriminating rand-big vs. info-big reward delivery (**Fig. S2b**, x-axis). Normally the model is inhibited by rand-big delivery and non-responsive to info-big delivery. But if the model is in the disengaged state then both rand-big or info-big deliveries would cause maximal inhibition, at least as large as rand-big delivery. So considering info-big trials, the model has a probability of $(1–p_d(IB))$ of being engaged and thus having a higher firing rate than all rand-big trials (ROC area as low as 0), whereas it has a probability of $p_d(IB)$ of being disengaged and thus having at least as low of a firing rate as all rand-big trials (ROC area no lower than 0.5). Hence the overall ROC area can be no lower than $(1–p_d(IB))*0 + p_d(IB)*0.5 = p_d(IB)*0.5$. By a similar argument there should also be a symmetrical upper bound on the ROC area, so that the ROC area can be no higher than $1–p_d(IB)*0.5$. Thus the ROC area should lie within a limited region

centered at 0.5 (**Fig. S2b**, extent of gray box along x-axis). By an analogous argument, we can also bound the ROC area for discriminating info-small from rand-small reward delivery (**Fig. S2b**, y-axis). That ROC area should be no lower than $p_d(R)*0.5$ and no higher than $1–p_d(R)*0.5$ (**Fig. S2b**, extent of gray box along y-axis).

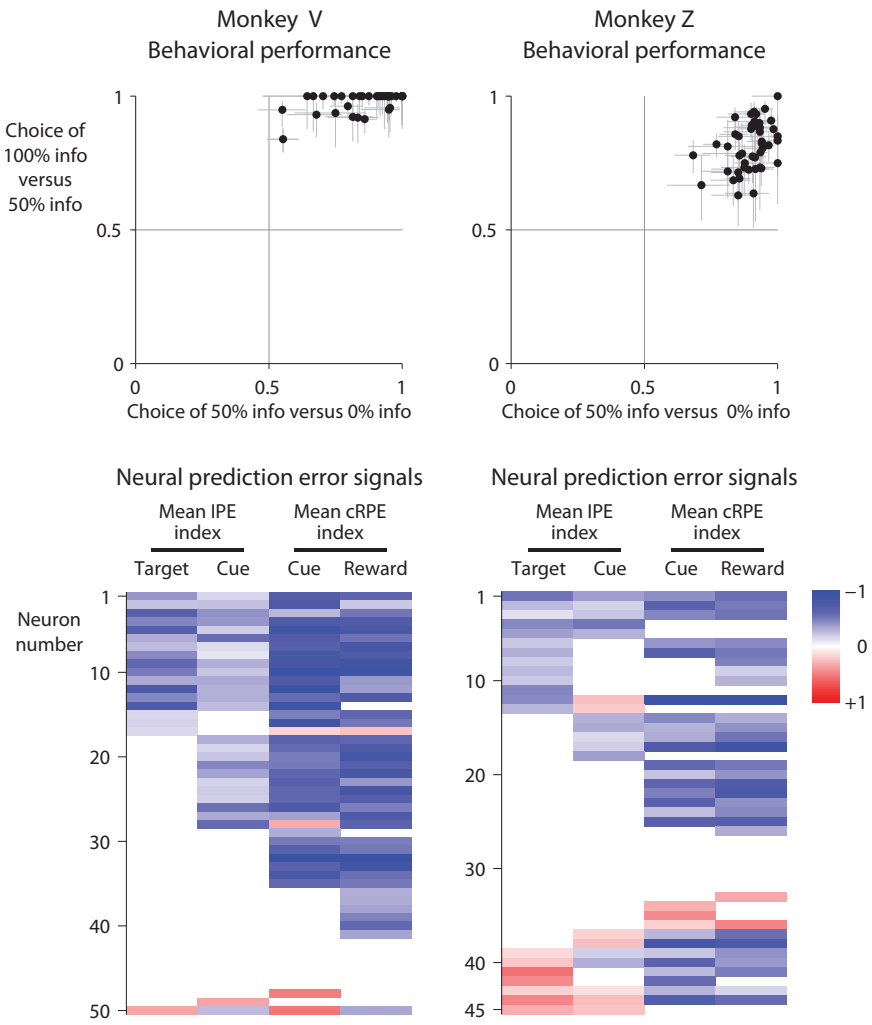
Results of disengagement model: (fitted parameters: $\beta = 21$, $k = –26$, $\varepsilon_0 = 0.60$, $\psi = 4.8$).

The disengagement model produced both IPE and cRPE signals similar to the information-bonus model and similar to those observed in the data (**Fig. S2a**, compare to **Fig. S1a,b**). The quality of the fit was slightly worse for the disengagement model despite its additional parameter (fitting error: 38 for the information bonus model, 53 for the disengagement model) but it produced the correct qualitative pattern of mean responses.

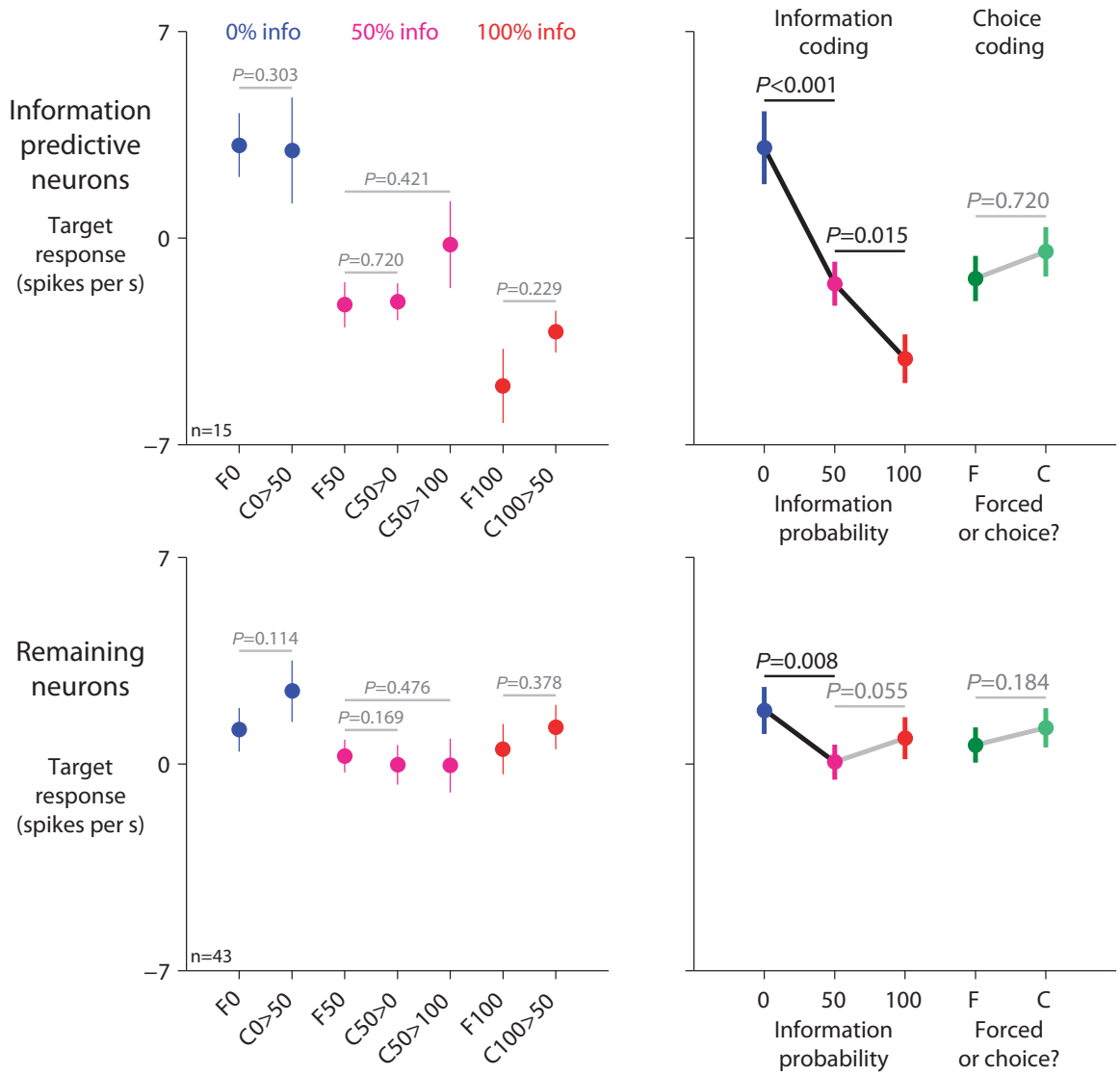
The disengagement model also predicted that the neural cRPE signals during reward delivery would have limited reliability, due to frequent disengagement that would cause the model to forget its reward predictions. According to the model the neural ROC areas for discriminating informed vs. random reward deliveries should be bounded within a limited range (**Fig. S2b**, gray box). In the data, however, most lateral habenula neurons (n=63/95) had ROC areas that fell outside that range (**Fig. S2b**, dots), indicating that the neurons had more reliable cRPE signals than predicted by the model.



Supplementary Figure 2. Test of computational mechanisms for assigning value to informative cues. (**a**) A modified TD model using a "disengagement" mechanism[43] produces qualitatively similar results to the information-bonus model (compare to Fig. S1b). It assigns greater value to informative cues than random cues, thus producing IPE signals and a behavioral preference to view the informative cues. (**b**) However, the disengagement model is inconsistent with the strength of neural discrimination between predictable and unpredictable reward delivery. The plot shows each neuron's ROC area for discriminating between predictable and unpredictable deliveries of big rewards (x-axis, Rand-big vs. Info-big) and small rewards (y-axis, Info-small vs. Rand-small). This is the same data as in Fig. 2b but expressed as ROC area. As expected, most neurons were clustered in the lower left corner of the plot, indicating strong coding of inverted cRPEs. According to the disengagement model, however, neurons should have been restricted to a limited region near the center of the plot (gray box, set according to the parameters of the best-fitting disengagement model), because frequent disengagement from the task would impair their ability to predict the reward size.



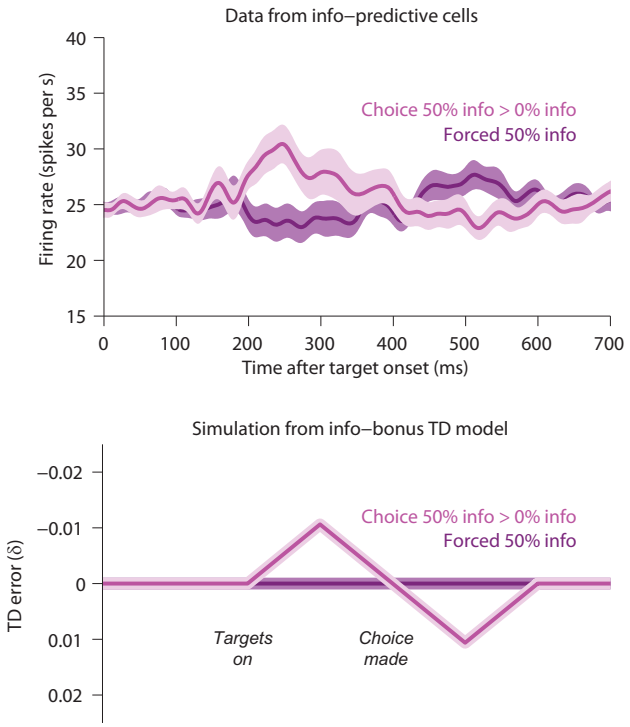
Supplementary Figure 3. Behavioral performance and neural activity for each recording session. Data are shown separately for monkey V (left) and monkey Z (right). Top: behavior. Fraction of trials when the animal chose 100% > 50% info (y-axis) or 50% > 0% info (x-axis). Error bars are ± 1 SE. Both monkeys expressed an orderly preference for 100% info > 50% info > 0% info during every recording session. Bottom: neural activity. Each row is a neuron. Colored patches indicate neurons with a significant mean IPE index for the target (first column), mean IPE index for the cue, mean cRPE index for the cue, or mean cRPE index for reward delivery (last column). Neurons are sorted based on whether they had significant effects in each successive column. The patch color indicates the direction of IPE/cRPE coding, with red for conventional prediction error signals (hypothesized for dopamine neurons) and blue for inverted signals (hypothesized for lateral habenula neurons). A considerable number of neurons had inverted IPE signals (blue, left two columns) and many of the same neurons also had inverted cRPE signals (blue, right two columns). A few neurons (primarily in monkey Z) had a tendency for inverted IPE or inverted cRPE signals (red patches). An example is Neuron D in Supplementary Figs. 9,10. We were unable to detect a significant session-to-session neural-behavioral correlation (mean percent choice of the higher info probability target vs. mean target IPE index, rho = -0.07, P = 0.51), perhaps because the animals assigned stable values to the targets so there was little underlying variability in target value.



Supplementary Figure 4. No significant difference in mean response to forced vs. choice trials. Cross-validated data are shown from the information-predictive neurons (top) and remaining neurons (bottom) that were recorded for at least one trial during all of the seven possible forced and choice trial conditions.

Left: Response to the targets for all seven trial types, sorted by the information probability of the chosen target (blue/purple/red indicate 0/50/100% info) and whether the trial was a forced or choice trial (text below x-axis indicates forced trials ("F") and choice trials ("C") as well as the non-chosen option (e.g. "C0>50")). There was no significant response difference between forced vs. choice trials that had the same information probability (gray text; all $P > 0.01$, signed-rank test).

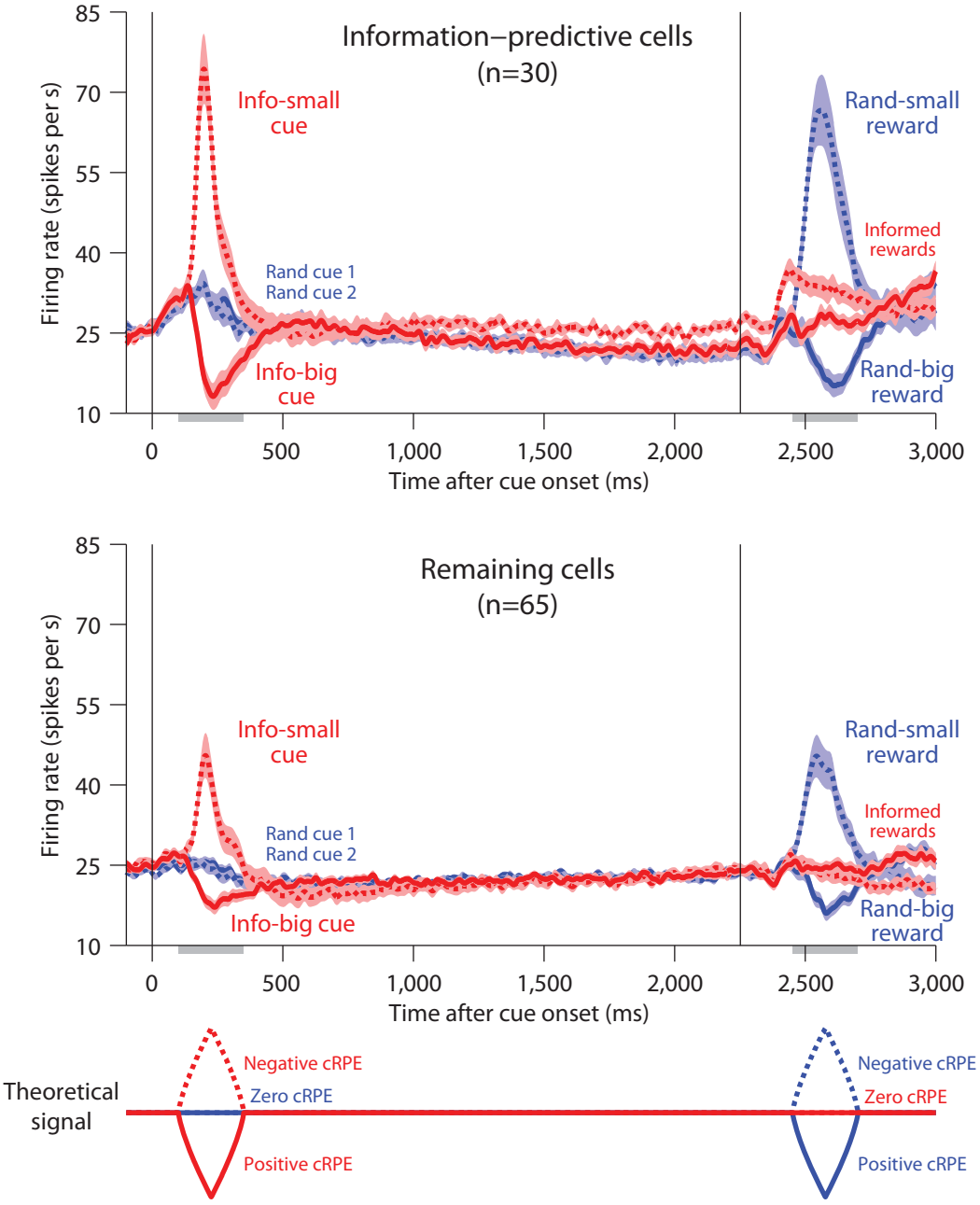
Right: Overall response magnitude coding information probability (blue/purple/red dots, mean of forced and choice conditions for each 0/50/100% info probability) and performance of forced vs. choice trials (dark/light green dots, mean of the 0/50/100% conditions for forced or choice trials). There were significant response differences based on information probability but not on forced vs. choice trials (text indicates p-values; signed-rank tests).



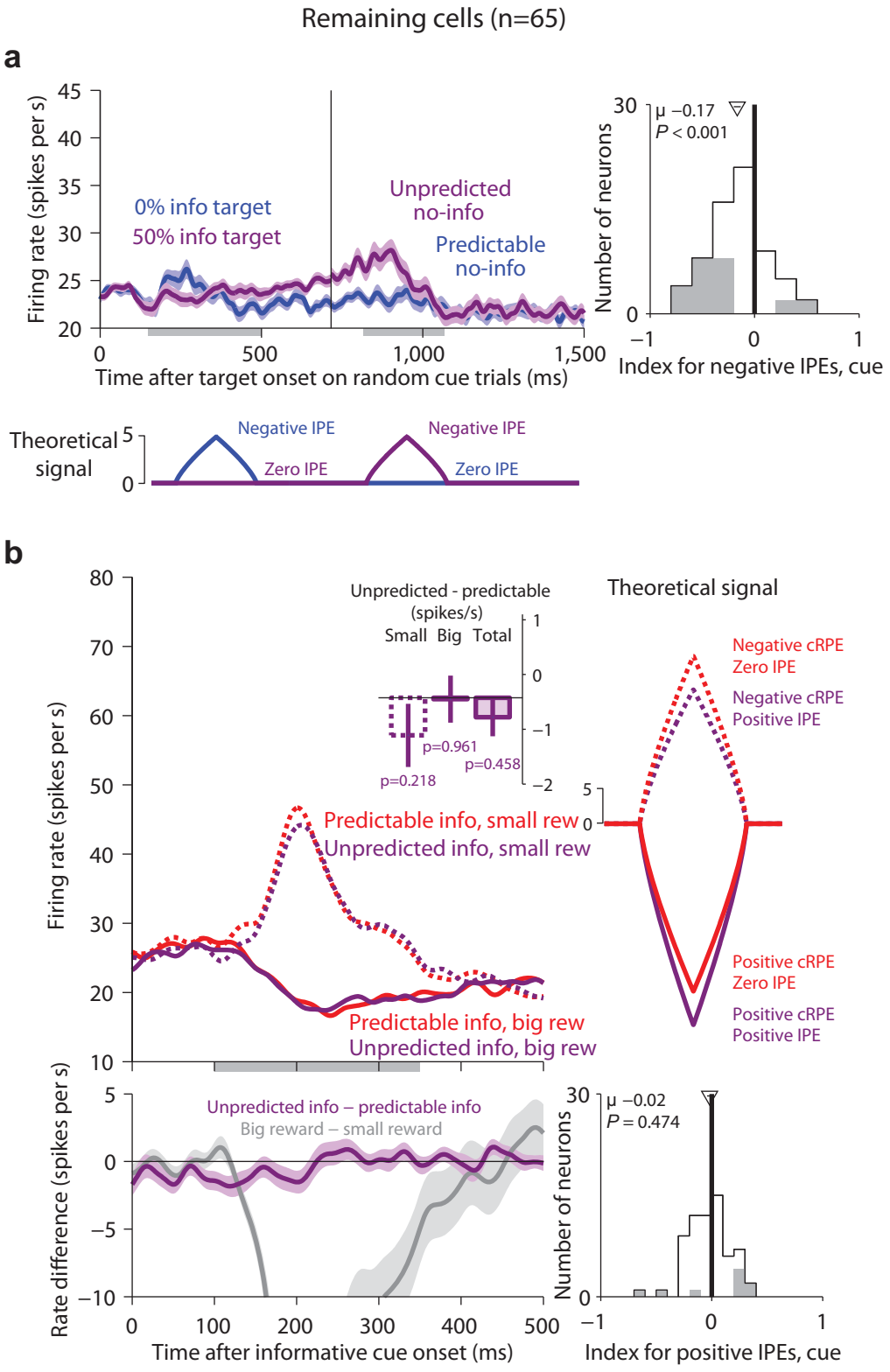
Supplementary Figure 5. Trend for different response time course on forced vs. choice trials. *Top*: cross-validated target responses from the information-predictive neurons on forced and choice 50% info trials (top, same as **Fig. 4a**). Although there was no significant difference in overall response magnitude on forced versus choice trials ($P = 0.11$, signed-rank test), there was a trend for different response time courses. Activity on choice trials tended to be initially higher, and later lower, than activity on forced trials.

Bottom: a similar biphasic trend can be seen in the TD error (δ) signals of the info-bonus TD model from Supplementary Fig. 1b, if its target response is decomposed into separate TD errors triggered sequentially by target array onset and choice onset (time axis: arbitrary units). This biphasic response occurs due to uncertainty about which of the targets the monkey will choose. On forced 50% info trials the monkey is certain to get the 50% info target, but on "50% vs. 0%" choice trials the monkey sometimes chooses the lower-value 0% info target, so the choice trials have a lower value than the forced trials. This causes two effects. First, it causes the choice array to evoke an initial small negative TD error (plotted here as an initial 'excitation' in the model for "Targets on"). Second, if the monkey then chooses the 50% info target after all, the value increases to become as high as on forced 50% trials causing a small positive TD error (plotted here as a later 'inhibition' in the model for "Choice made").

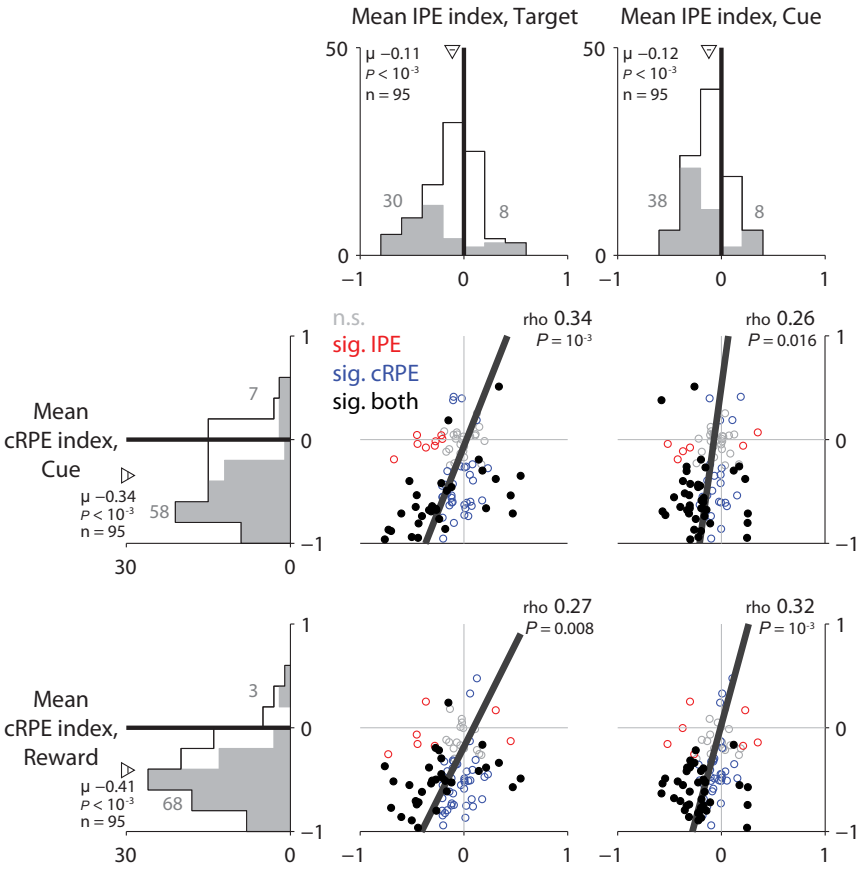
Note, however, that the neural data and model simulation are difficult to compare directly, for three reasons: (1) even if the two hypothesized TD errors occur, they may be intermixed instead of sequential; (2) the two resulting neural responses could have different durations and could overlap in time, in a manner that is difficult to predict; (3) based on the model parameters, each 'phase' of the model response represents a relatively small change in spike count (about 0.1 spikes per trial) but could translate into either a large or small change in spike rate, depending on its duration.



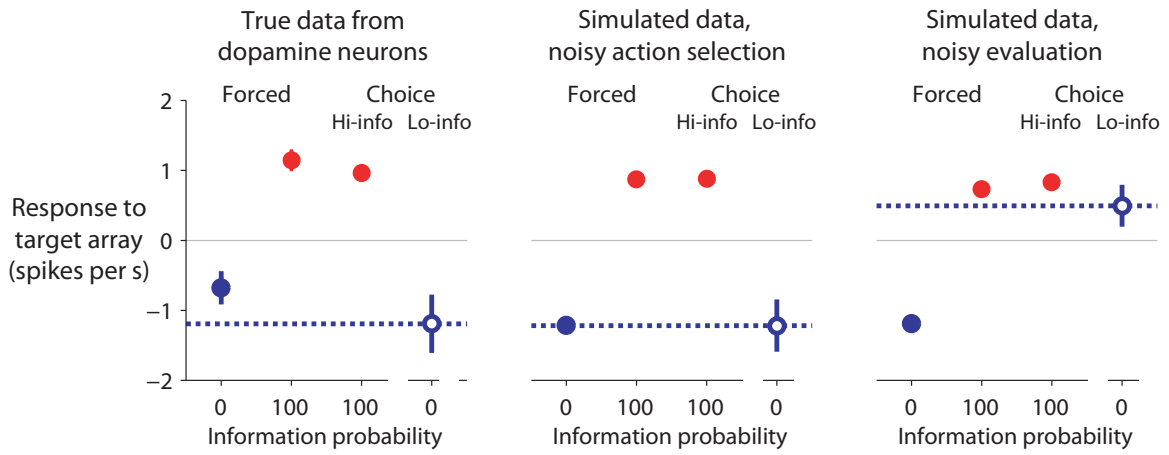
Supplementary Figure 6. Conventional RPE signals in the information-predictive and remaining neurons. Same as Fig. 2, shown separately for the information-predictive neurons (top) and the remaining neurons (bottom). Both subpopulations of neurons had similar reward prediction error signals, although these signals tended to be stronger in the information-predictive neurons.


Remaining cells (n=65)
a
Firing rate (spikes per s)
45
40
35
30
25
20
0% info target
50% info target
Unpredicted
no-info
Predictable
no-info
0
500
1,000
1,500
Time after target onset on random cue trials (ms)
Number of neurons
30
0
μ −0.17
P < 0.001
−1
0
1
Index for negative IPEs, cue
Theoretical signal
5
0
Negative IPE
Zero IPE
Negative IPE
Zero IPE
b
Firing rate (spikes per s)
80
70
60
50
40
30
20
10
Unpredicted - predictable
(spikes/s)
Small Big Total
1
0
−1
−2
p=0.961
p=0.458
p=0.218
Predictable info, small rew
Unpredicted info, small rew
Predictable info, big rew
Unpredicted info, big rew
Theoretical signal
Negative cRPE
Zero IPE
Negative cRPE
Positive IPE
5
0
Positive cRPE
Zero IPE
Positive cRPE
Positive IPE
Rate difference (spikes per s)
5
0
−5
−10
Unpredicted info – predictable info
Big reward – small reward
0
100
200
300
400
500
Time after informative cue onset (ms)
Number of neurons
30
0
μ −0.02
P = 0.474
−1
0
1
Index for positive IPEs, cue

Supplementary Figure 7. Absence of consistent IPE signals in the remaining neurons. (**a**) Same format as Fig. 5a,b, but showing activity from the remaining neurons. Similar to the information-predictive neurons, many of the remaining neurons had higher activity in response to 'unexpected no-info' (purple, random cues after 50% info target) than in response to 'expected no-info' (blue, random cues after 0% info target). Thus, their median negative IPE index for the cue was less than zero (-0.17; signed-rank test, $P < 0.001$). This activity was somewhat different from the theoretical inverted IPE signal, however (bottom), because it was a sustained change in activity that occurred even before the cues were presented, perhaps reflecting a long-latency neural response to the 50% and 0% info targets. (**b**) Same format as Fig. 6a-c, but showing activity for the remaining neurons. Unlike the information-predictive neurons, the remaining neurons had smaller and non-significant differences in their activity between 'unpredicted info' trials (purple) and 'predictable info' trials (red) (inset bar plot: all differences < 1 spike/s, all $P > 0.2$, signed-rank test; bottom plot of differences in firing rate: no stimulus-triggered difference in neural activity). Thus, their activity was different from the theoretical inverted IPE+cRPE signal (right), and they had a mean positive cue IPE index of just -0.02, not significantly different from zero ($P = 0.47$, signed-rank test).



Supplementary Figure 8. A further test of IPE and cRPE transmission by single neurons. For each neuron we calculated the mean IPE index and mean cRPE index for each task event. Marginal histograms show the single neuron distribution of the mean cRPE indexes for the visual cues and for reward delivery (left, y-axis of scatterplots), and the mean IPE indexes for the target array and for the visual cues (top, x-axis of scatterplots). Gray numbers indicate count of neurons have mean cRPE or IPE indexes significantly different from 0 (gray neurons on histograms; $P < 0.05$, permutation test). Text indicates the average of the single neuron indexes and the p-value (signed-rank test). Scatterplots show the relationship between the single neuron mean cRPE indexes and mean IPE indexes. Colors indicate significance of the indexes ($P < 0.05$, permutation test), showing neurons with no significant indexes (gray), a significant IPE index (red), a significant cRPE index (blue), or both significant indexes (black filled circles). Text indicates rank correlation (rho) and its p-value (permutation test); sold line indicates best-fitting linear relationship using type 2 regression. This analysis indicated that IPE and cRPE signals often occurred in the same neurons (black cells in lower left quadrant of each plot). For example, the 30 information-predictive neurons had significant inverted IPE coding for the target (mean target IPE index < 0; permutation test, $P < 0.05$), and 20/30 also had significant inverted cRPE coding for both the reward cues and reward delivery (both mean cRPE indexes < 0, $P < 0.05$). A total of 15 neurons had significant inverted prediction error coding for all four task events – IPE signals for the target array and the cues, and cRPE signals for the reward cues and deliveries. Furthermore, the indexes were correlated so that cells with strong IPE coding also tended to have strong cRPE coding (all rho > 0.25, all $P < 0.02$; permutation tests).



Supplementary Fig. 9. Neural data are consistent with low-info choices being caused by noisy action selection. We compared the true neural responses (left, same as **Fig. 8f**) with simulated neural responses representing the hypotheses that low-information choices are caused by *noisy action selection* (middle panel) or *noisy evaluation* (right panel). For simplicity, we considered the experiment used to record dopamine neurons in which there were only two targets (100% and 0% info) that were presented in a mixture of both forced and choice trials (Methods). In the simulations the mean target values were set equal to V(100% info) = 2 and V(0% info) = 0. The simulated dopamine neurons signaled prediction errors defined as the value of the chosen option minus the expected value of the trial (which was computed based on the values of both potential options). The simulations were adjusted to resemble the neural recording sessions in terms of the probability of choosing the low-info target (~10%), the number of trials (n=3000), and the trial-to-trial variability in firing rates.

In the *noisy action selection* simulation, the values of the two targets were constant on all trials but the animal had a 10% probability of making an error in action selection and choosing the low-value target. This produced a pattern of results closely resembling the true neural activity observed in dopamine neurons (compare left panel to middle panel).

In the *noisy evaluation* simulation, the animal always chose the target that had a higher value but the values were perturbed from trial to trial with random Gaussian noise to induce a 10% probability of choosing the low-info target. This produced a very different pattern of results: prediction errors were much more positive during choice 0% info trials (blue open circle, dotted line) than during forced 0% info trials (blue filled circle). This is different from the pattern seen in the neural data (compare left panel to right panel). This occurred because the simulated animal only chose the low-info target on trials when evaluation noise gave it a high perceived value, so that low-info choices were associated with overly optimistic prediction errors.

Full description for Supplementary Figs. 10,11
IPE and cRPE signals in single neurons – summary and raw activity

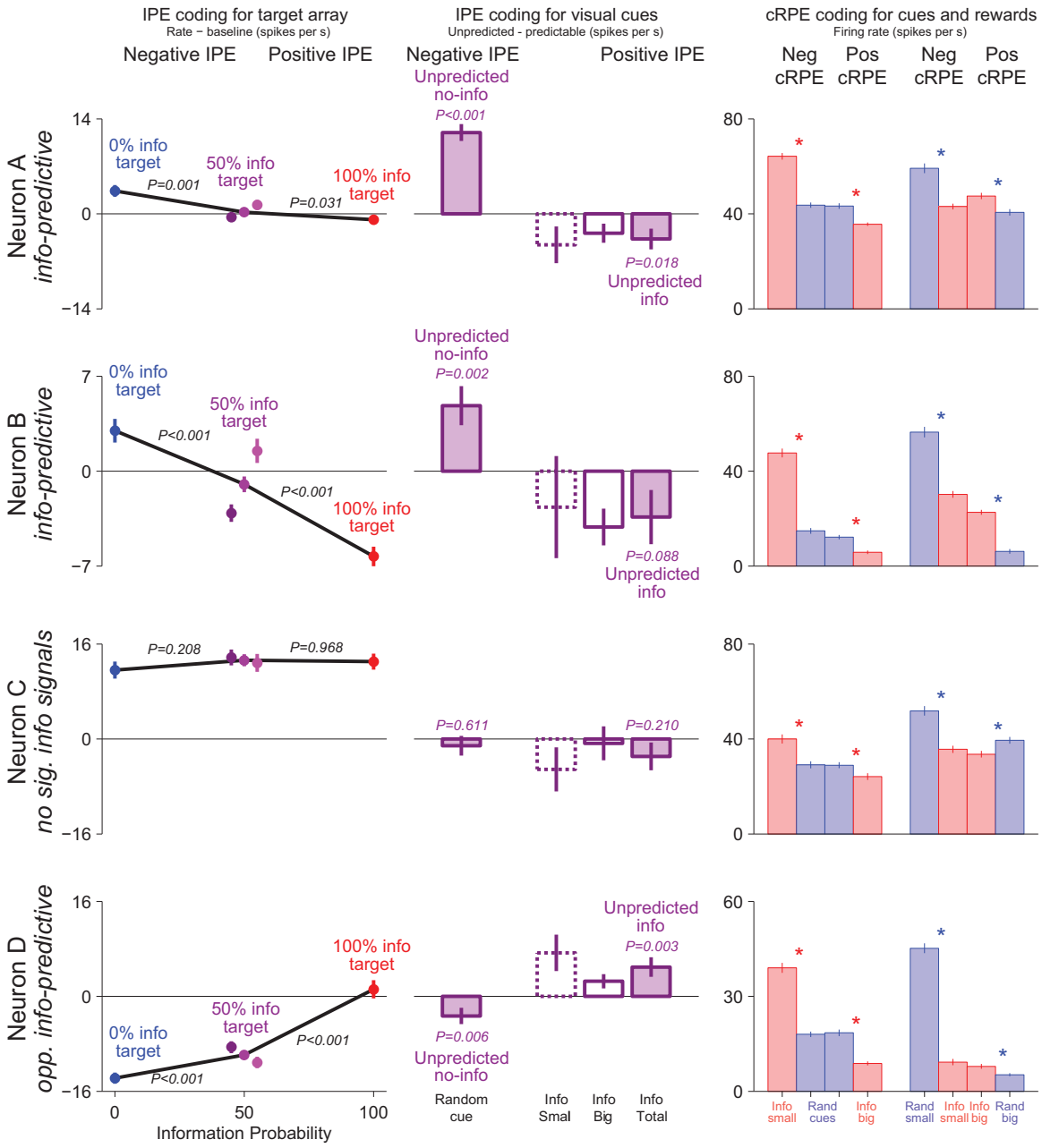
Our analysis in the main text shows that a considerable subpopulation of lateral habenula neurons transmitted signals related to IPEs, and most also transmitted signals related to cRPEs. This suggests that downstream regions could decode lateral habenula cRPE and IPE signals by simply averaging spikes across many lateral habenula neurons. At the level of single neurons, however, we observed that these IPE and cRPE signals were often superimposed on additional heterogeneous forms of tonic and phasic activation. This suggests that downstream regions might be able to decode additional information about task events by considering the activity levels of individual lateral habenula neurons. To illustrate, we have plotted data from four example neurons chosen to show the most common forms of IPE/cRPE and non-IPE/cRPE related activity in single cells (**Figs. S10,11**). The first plot shows a summary analysis of each neuron's IPE and cRPE signals based on comparisons of neural firing rates (**Fig. S10**). The second plot shows each neuron's raw activity, to illustrate their other forms of task-related activity (**Fig. S11**). A detailed description of their activity is below.

The first two examples are information-predictive neurons (neurons A, B). These neurons had inverted IPE signals in response to the targets and cues, as seen in the population as a whole (**Figs. 4-6**). Their target responses were inversely related to information probability (left column); they were excited during 'unpredicted no-info' (middle column); and they tended to be relatively inhibited during 'unexpected info' (middle column). These IPE signals were superimposed on additional heterogeneous tonic and phasic activations. These included target responses with multiple excitatory and inhibitory phases which could occur differently during forced vs. choice trials (neuron A, compare forced 50% vs. choice 50% trials; forced trials have a biphasic inhibition-excitation, while choice trials have a triphasic inhibition-excitation-inhibition). These neurons also had cRPE signals similar to those seen in the population as a whole (right column). Again, these cRPE signals could be superimposed on additional heterogeneous tonic and phasic activations. After negative cRPEs triggered by the info-small cue, some neurons had sustained tonic inhibition (neuron A) or tonic excitation (neuron B) lasting throughout the cue period. After delivery of a big reward, some neurons had post-outcome tonic excitation (neuron B, and to a lesser extent neuron A; right column, far right side of plot, solid lines above dotted lines), as reported previously[44].

Neuron C is an example that had cRPE-like signals but no significant IPE signals – one of the "remaining neurons" described in the main text (**Fig. 4**). This neuron did not seem to encode IPEs because it responded to the targets with a mostly non-differential excitation (left column), and it had no significant response modulation by 'unpredicted no-info' or 'unpredicted info' (middle column). But this neuron had an cRPE-like response to the cues (right column: excitation by info-small cue, inhibition by info-big cue, intermediate response to random cues) and to reward delivery (right column: initial non-differential excitation followed by further excitation by rand-small rewards and relative inhibition by rand-big rewards). Again, these cRPE-like signals were superimposed on other heterogeneous forms of task-related activity, such as an overall suppression of firing rate during the cue period on trials when the info-big or random cues were presented (right column, compare blue and solid red lines to red dashed lines).

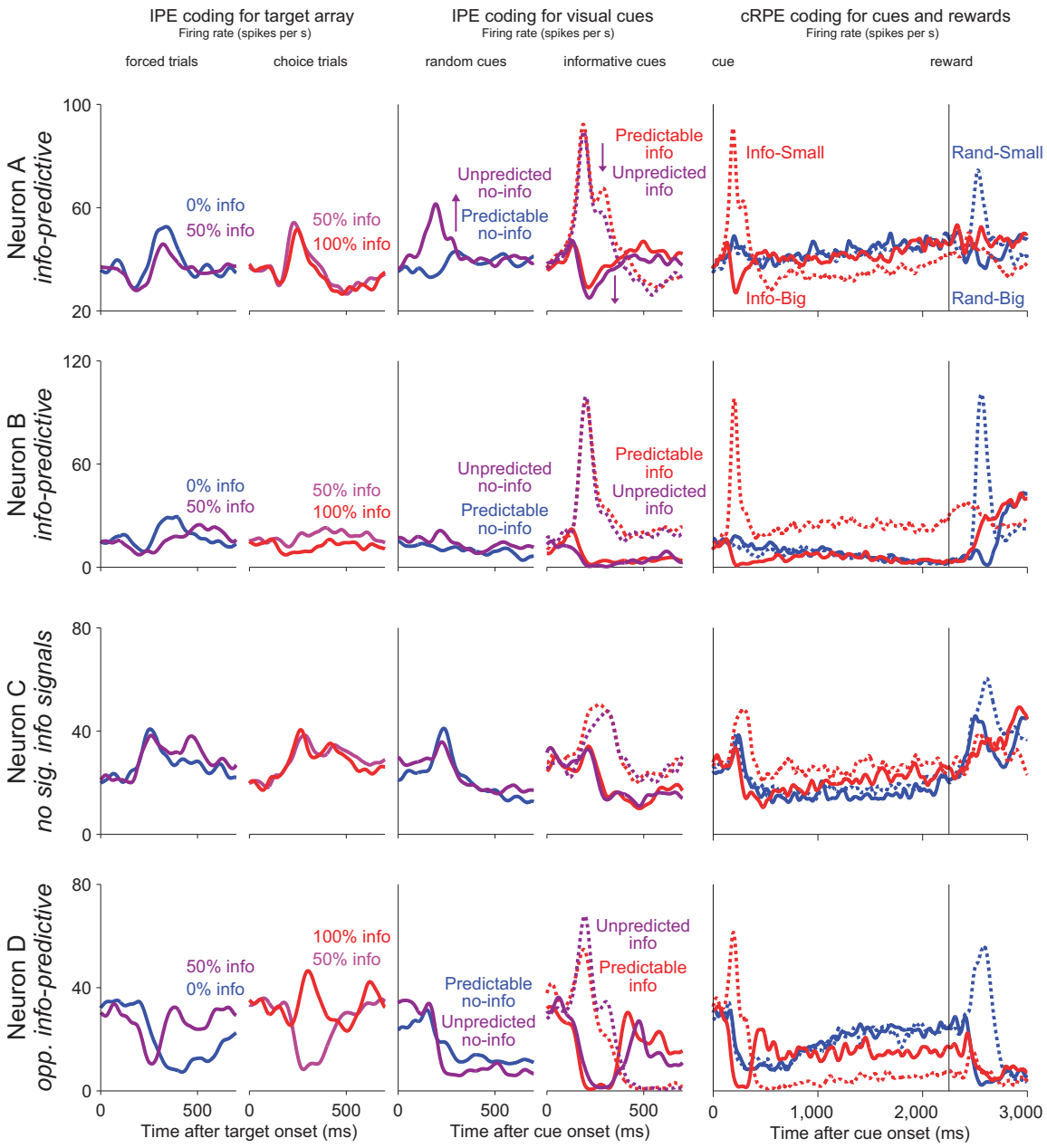
Neuron D was a rare example that had IPE signals in the opposite direction of the lateral habenula population as a whole. This neuron's target response was positively related to information probability (left column). Its random cue responses were relatively inhibited by

negative IPEs (‘unpredicted no-info’ < ‘predictable no-info’, middle column) and its informative cue responses were relatively excited by positive IPEs (‘unpredicted info’ > ‘predictable info’; middle column). Despite this unusual direction of IPE coding, this neuron still had the typical habenular direction of phasic cRPE-like signals in response to the cues and outcomes (right column). It was excited by negative IPEs (right, info-small cue and rand-small reward) and inhibited by positive IPEs (right, info-big cue and rand-small outcome). These IPE and cRPE signals were again superimposed on other heterogeneous forms of activity such as an overall suppression of firing rate during the cue period when the info-small cue was presented (right column, red dashed line below blue lines and red solid line) and an overall suppression of activity after reward delivery (right column, far right side of plot).



Supplementary Figure 10. Signals related to IPEs and cRPEs in four example neurons – summary. Each row is one neuron. **First column**: IPE coding in response to the target array (same format as Fig. 3C). Plotted is the baseline-subtracted firing rate. Text is p-value of the difference in firing rate for negative IPEs (forced 0% info vs. forced 50% info, left two dots) and positive IPEs (choice 50% info vs. choice 100% info, right two dots). Error bars are ±1 SE. **Second column**: IPE coding in response to the visual cues, separately for negative IPEs (random cues, 'unpredicted no-info' – 'predictable no-info', left, purple bar) and positive IPEs (informative cues, 'unpredicted info' – 'predictable info', right; data shown for small-reward cue (dashed white bar), big-reward cue (solid white bar), and average of both cues (purple bar)). Error bars are ±1 SE. Text indicates p-value of the difference in firing rate between unpredicted vs. predictable trials. **Third column**: cRPE coding in response to the cues and rewards. Plotted

are the firing rates in response to the cues (left: info-small cue, random cues, and info-big cue) and rewards (right: random small, informed small, informed big, and random big). Asterisks indicate significant differences in rate for negative cRPEs (difference between info-small vs. random cue, or between rand-small vs. informed-small reward) and positive cRPEs (difference between info-big vs. random cue, or between rand-big vs. informed-big reward). See associated text for a detailed description of each neuron.



Supplementary Figure 11. Signals related to IPEs and cRPEs in four example neurons – raw activity. Each row shows activity from a single neuron, smoothed with a Gaussian kernel (σ=20 ms). **First column**: IPE coding in response to the target array. Same format as Fig. 3a,b. Data are shown separately for negative IPEs (left, forced 0% vs. forced 50% info) and positive IPEs (right, choice 100% vs. choice 50% info). **Second column**: IPE coding in response to the visual cues. Similar format to Fig. 5a and 6a. Data are shown separately for negative IPEs (left, random cues, 'unpredicted no-info' vs. 'predictable no-info') and positive IPEs (right, informative cues, 'unpredicted info' vs. 'predictable info'). **Third column**: cRPE coding in response to the cues and rewards. Same format as Fig. 2a. See associated text for a detailed description of each neuron.

Supplementary References

51. Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci* **16**, 1936-47 (1996).
52. Sutton, R. S. & Barto, A. G. Reinforcement learning: an introduction (MIT Press, 1998).