Network computations underlying learning from symbolic gains and losses

Hua Tang[1*], Ramon Bartolo[1, 2], Bruno B. Averbeck[1,3*]

1. Laboratory of Neuropsychology, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA.

2. Laboratory of Sensorimotor Research, National Eye Institute, National Institutes of Health, Bethesda, MD, USA.

3. Lead Contact

*Correspondence: hua.tang@nih.gov (H.T.), bruno.averbeck@nih.gov (B. B. A.)

SUMMARY

Reinforcement learning (RL) engages a network of areas, including the orbitofrontal cortex (OFC), ventral striatum (VS), amygdala (AMY), and mediodorsal thalamus (MDt). This study examined RL mediated by gains and losses of symbolic reinforcers across this network. Monkeys learned to select options that led to gaining tokens and avoid options that led to losing tokens. Tokens were cashed out for juice rewards periodically. We found that task-relevant information was distributed across the network. However, examination of the way in which information was encoded differed, with VS showing increased responses to appetitive outcomes, OFC differentiating primary and symbolic reinforcers, and AMY responding to the salience of outcomes. In addition, analysis of network activity showed that symbolic reinforcement was calculated by temporal differentiation of accumulated tokens. This process was mediated by dynamics within the OFC-MDt-VS circuit. Thus, we provide a neurocomputational account of learning from symbolic gains and losses.

INTRODUCTION

Neurophysiology and anatomy have identified large-scale forebrain networks that underlie specific behavioral processes [1,2]. Most of the work in physiology has focused on establishing the function of single cortical or subcortical areas, by looking for correlates of hypothesized functions in the responses of single neurons or populations within each area. The recent advent of circuit work has extended the study of single areas to pairs of areas, by looking at specific inputs to areas or projection targets of areas [3]. However, this work has not extended the study to networks of areas. Much of this work has shown that anatomically connected areas tend to process similar information once one moves beyond primary sensory or motor areas [4], and lesions or manipulations to multiple nodes across a connected network can lead to similar deficits in tasks [5,6]. An exception to this is fMRI, which has shifted to studying the contribution of identified networks to behavior [7,8]. Thus, examining connected areas within networks is important to understand how the network orchestrates behavior.

Although early studies of reinforcement learning (RL) focused on the striatum and the dopamine innervation of the striatum [9], subsequent work showed that additional structures, including a network composed of the OFC [10,11], AMY [12,13], and MDt [5] were also involved in RL [2]. Although causal manipulations and neurophysiology have shown that each of these areas plays a role in RL, the results across tasks and areas are not always consistent. For example, lesions to the VS cause deficits in bandit tasks that use probabilistic delivery of primary reinforcers [13]. However, these deficits are limited to learning to associate rewards with objects and not actions [14]. This is not consistent with actor-critic models of RL that suggest that VS stores a general state value representation for policy learning [15,16]. Similarly, AMY and VS appear to play stronger roles in probabilistic bandit tasks but not in token-based RL tasks, in which tokens are used as reinforcers [17-19]. In token-based RL, VS is only involved in learning to discriminate between relative gains and plays no role in learning to discriminate between gains and losses [17]. AMY, on the other hand, appears to play almost no role in token-based RL [18]. Previous neurophysiology studies have shown that token-based learning may engage cortical networks more than subcortical networks [20,21], whereas learning based directly on primary rewards may preferentially engage subcortical networks [22,23]. Beyond these examples, some learning-related behaviors may rely more on inference processes than incremental value updates that characterize RL and may, therefore, tap into different networks [24,25]. Thus, the way in which the ventral cortico-striatal-thalamo-cortical network, including the OFC, VS, AMY, and MDt, orchestrates RL, particularly with symbolic reinforcers, is unclear.

In the present study, we have carried out physiology recordings across the ventral cortico-striatal-thalamo-cortical network using a token-based RL task. We examined single neuron, population, and network

computations underlying the performance of the task. We began by examining single neuron coding of task variables across areas, and found, consistent with previous work, that the task-relevant variables were encoded broadly across areas. We found a small enhancement of token encoding in the OFC, chosen stimulus identity in the AMY, and choice direction (action) encoding in the MDt. However, other variables, including chosen value and primary reward, were encoded similarly across areas. Thus, finding a neural correlate of a task variable within a single area provides minimal evidence that the area is important for that variable. More differences emerged when we examined how neurons were tuned to task variables across areas. Specifically, although both OFC and VS encoded chosen value and token outcomes, VS strongly encoded appetitive choices and outcomes (i.e., gains) with increased firing rates, while OFC was more balanced with a bias towards encoding appetitive choices and outcomes with decreased firing rates. AMY uniquely showed enhanced salience coding and no correlation between the encoding of chosen value and token outcomes. Furthermore, AMY showed the strongest correlations between population coding of primary and symbolic reinforcers, whereas OFC strongly encoded both reinforcers but discriminated between them at the population level. Finally, when we examined network computations underlying token reinforcement, we found that OFC and VS both computed token outcomes as time derivatives of accumulated tokens and relayed this information reciprocally through the MDt. Taken together, this shows unique contributions of each area, revealed by analysis of large-scale recordings across the ventral cortico-striatal-thalamo-cortical network, as well as evidence for why lesions across the nodes can affect RL differentially with different paradigms.

RESULTS

Two rhesus monkeys were trained on a two-armed bandit task. In this task, the monkeys collected tokens, which were periodically exchanged for juice rewards (Figures 1A-B). In every block of 108 trials, we introduced four novel images. In each trial, two of the four images were presented on the screen. The choice of an image led stochastically to one of -2, -1, +1, or +2 tokens (Figure 1B). The image-outcome relationships were unknown to the monkeys at the start of the block. Monkeys had to learn the values of the images by choosing one of them and observing the outcome. Token outcomes were stochastic such that in 75% of the trials, the monkeys received the number of tokens associated with the chosen image, and in 25% of the trials, the number of tokens did not change. Tokens were accumulated across trials and were cashed out every four to six trials, with one drop of juice for each token.

The primary reward was apple juice (#juice), but it was only delivered every four to six trials. In each trial, the monkeys had a chance to gain or lose tokens by choosing one of the images. The chosen value information was first carried by the images (cValue), which predicted the change of tokens (Δtoken). The number of accumulated tokens (#token) was always presented on the screen. So, value information appeared in different forms, including cValue, Δtoken, #token, and #juice (Figure S1A). The following analyses focus on the neural representation and interactions of these forms of value.

Choice behavior was influenced by gaining/losing tokens and token numbers

We quantified choice behavior by measuring the fraction of times the monkeys chose the image associated with a higher value in each condition (e.g., choose +1 when presented with -1 and +1 images). The monkeys learned to distinguish the values of two images shown on the screen within about 10 trials for the Gain/Loss and Gain/Gain conditions but learned minimally in the Loss/Loss condition (Figures 1G, S1B). We also fit a Rescorla-Wagner (RW) reinforcement learning model to the choice behavior (Figure 1G), which has been explored previously [17]. The model is used below to examine token reward prediction error (RPE). To quantify how the choice behavior (i.e., whether they chose the better option) was affected by task variables, we fit a multi-way ANOVA model to it (Figures 1H-I). The choice behavior was significantly modulated by the number of observations (Trial, $F_{17, 44950} = 109.98$, $p < 0.001$), Gain/Loss condition (Gain/Loss, $F_{1, 44950} = 286.91$, $p < 0.001$), value difference of two options (Δvalue, $F_{3, 44950} = 73.37$, $p < 0.001$), and the number of tokens before choice (#token, $F_{10, 44950} = 8.68$, $p < 0.001$), but not by the number of trials since last token cashout (CashID, $F_{5, 44950} = 2.07$, $p = 0.066$). The number of observations indicates the learning process, both the Gain/Loss condition and Δvalue indicate the updating of tokens, #token indicates the

accumulated value, and the CashID indicates the probability of receiving a primary reward. The result indicates that the monkeys understood the meaning of tokens and adjusted their behavior to get more tokens as learning progressed.

Neural encoding of choices, primary and symbolic reinforcers

We performed simultaneous recordings of population neural activity across four regions of the ventral cortico-striatal-thalamo-cortical network (Figures 1C, S1E) from two macaque monkeys using multi-site linear probes. We collected 606 neurons in the 13L region of the orbitofrontal cortex (OFC), 829 neurons in the core region of the ventral striatum (VS), 1607 neurons in the basolateral amygdala (AMY), and 1035 neurons in the medial portion of the mediodorsal thalamus (MDt) using a semi-chronic recording procedure (Figures 1D-F). Neuronal activity differed across areas in different task epochs (Figures S1F-G).

To examine how choices and rewards were represented in each area, we fit the responses of single neurons with a sliding window ANOVA model. The model included multiple task-relevant factors, which were the number of tokens (#token), the change of token numbers (Δtoken), the number of juice drops delivered on cashout trials (#juice), the image pair presented (condition), the stimulus identity (cStim), *a priori* value (cValue, i.e., +2, +1, -1, -2), and direction (cDir) of chosen images. The factors #token, Δtoken, and #juice (Figures 2A-C) were the reinforcement signals that drove the monkeys' behavior. Neurons in all areas (> 10%), but more in the OFC (> 20%), showed a strong representation of #token (Figures 2A, H). This is consistent with the tokens that were always present on the screen during the trial. There was also a phasic increase in all areas when tokens were updated. The Δtoken was also encoded by neurons in all areas but more strongly in the OFC and VS (Figure 2B). OFC also led the encoding of Δtoken in time (Figure 2B, inset). The #juice activated more than 40% of neurons in every area (Figures 2C, I). OFC played a significant role in encoding the reinforcement signals, having the highest proportion of neurons encoding these task variables. A substantial proportion of neurons in these areas also showed responses to choices, including the identity (Figure 2D), *a priori* value (Figures 2E, G), and direction (Figure 2F) of chosen images. Learning-related values derived from the RW model were encoded consistently across areas (Figure S4B). The stimulus identities, chosen values, and directions were most strongly represented in the AMY, OFC, and MDt, but differences were often subtle. This result shows that task variables related to the choices and rewards were encoded across the ventral network, with different regions encoding specific types of information at different phases.

Diverse encoding of value update information

The updating of value information was represented as gaining or losing tokens. In the first analysis, we used only linear encoding (Figure 2B). However, neurons may encode value information in more diverse ways. To quantitatively characterize the encoding patterns across the areas, we examined the activity of each neuron using a series of linear regression models (Figure S2A). We identified the best-fitting model for each neuron and classified them into different functional categories according to how they were tuned to outcomes.

Here, we classified neurons encoding Δtoken into five types: neurons encoding (1) value linearly across both Gain and Loss (Figures 3A, F), (2) value salience (Figures 3B, G), (3) categorial Gain/Loss (Figures 3C, H), (4) value only for Gain (Figures 3D, I), and (5) value only for Loss (Figures 3E, J). Note that each encoding type could be positively or negatively related to the neural activity. The first category encoded Δtoken linearly (Figure 3A). These neurons encoded the gains and losses on a linear value axis, in other words, processing the gaining and losing of tokens in the same internal value system. Many OFC neurons, but few in the AMY, were of this type. However, more AMY neurons encoded the salience of Δtoken. These neurons represent both gains and losses but with inverse correlations of neural activity and value (Figure 3B). The categorical Gain/Loss signal groups each option as gain or loss, regardless of the magnitude. These neurons were found throughout the ventral network, especially in the OFC and VS (Figure 3C). Many more neurons in these areas encoded gains rather than losses. VS and OFC neurons showed robust and phasic responses to gains (Figure 3D). Only a small proportion of neurons encoded losses, mostly in the OFC (Figure 3E). Before the token update, the chosen images predicted the value update. We also classified the neurons encoding cValue into the same five types (Figure S3A). They showed similar encoding patterns, with activity locked to the onset of the images. This analysis shows that these areas played unique roles in encoding the chosen value and token outcome information.

Populational representation of gains and losses

An alternative way to examine the encoding of value update information is to measure the representations of gains and losses separately by running two regressions, either using Gain trials (i.e., $\Delta token \in [0, 1, 2]$) or Loss trials (i.e., $\Delta token \in [0, -1, -2]$), then calculate the correlation of the Gain and Loss regression coefficients (Figures 3K-O). The OFC, VS, and MDt populations showed high co-encoding of gains and losses, which indicates the encoding of gains and losses on similar value axes (Figure 3K). Especially for OFC, the effect was consistent for two seconds after the token update (Figure 4L). This was consistent with the single-cell result (Figure 3A), indicating that OFC processes the gaining and losing of tokens closer to

163 objective outcomes than the other areas (Figure 4K; Fisher's z-transformation, $p < 0.05$). Conversely, AMY
164 showed negative correlations between Gain and Loss regression coefficients (Figures 4K, N), indicating
165 populational encoding of value salience (Figure 3B).

166 Co-encoding of the value update information

167 In each trial, the monkeys may gain or lose tokens. The predicted value update was carried by the images
168 (cValue) before the token outcome was delivered (Δtoken). Did neurons in these areas encode the two forms
169 of value information in a similar way? To address this question, we first compared the number of neurons
170 that encoded both forms of value linearly. Neurons encoding cValue in a window 500 ms before the token
171 update and Δtoken in a window 500 ms after the token update were classified as encoding both cValue and
172 Δtoken (blue dots in Figures 4A-D). Fewer neurons in the AMY than in the other areas linearly encoded
173 both factors (Figures 4A-D, F; chi-square test, $p < 0.01$). To address how similar AMY neurons encoded
174 cValue and Δtoken, we then calculated the correlation of the cValue and Δtoken regression coefficients in
175 each area (Figures S2B-C). AMY neurons showed a significantly lower correlation, r, between cValue and
176 Δtoken than the other areas (Figures 4A-D, G; Fisher's z-transformation, $p < 0.05$ for significant neurons,
177 $p < 0.001$ for all neurons). This result shows that neurons in the AMY encoded the value carried by the
178 images and token outcome less consistently, which is also consistent with AMY primarily encoding
179 salience.

180 We also found asymmetric encoding of gains and losses in the VS, OFC and MDt (Figure 4E; chi-square
181 test; VS, $x^2 = 95.74$, $p < 0.001$; OFC, $x^2 = 19.96$, $p < 0.001$; MDt, $x^2 = 10.87$, $p < 0.05$). More VS neurons
182 had positive cValue and Δtoken regression coefficients. Together with the results in Figure 3, this indicates
183 that more neurons in the VS fired more when choosing higher-valued images and getting more tokens in
184 the Gain conditions (Figures 4B, E, quadrant 4). OFC neurons showed more balanced encoding in increased
185 and decreased activities (Figure 4E, quadrant 3 vs. 4; chi-square test, $x^2 = 3.11$, $p = 0.078$). Thus, neurons
186 in the VS tended to respond with increased firing rates to the choice of good options and outcomes, whereas
187 neurons in the OFC had both positive and negative tuning to the same variables.

188 Co-encoding of primary and symbolic reinforcers

189 Although a growing number of studies have adopted symbolic reinforcers, whether they are encoded in the
190 same way as primary reinforcers remains an open question. To address this, we compared the encoding of
191 #token and #juice. We first calculated the number of neurons encoding each variable. Neurons encoding
192 #token in a window 500 ms after the token update and encoding #juice in a window 500 ms after juice

193 delivery were classified as encoding both #token and #juice (Figure S3B). More neurons in the OFC than
194 in the other areas encoded both #token and #juice (Figure 4H; chi-square test, $p < 0.001$). We then calculated
195 the correlation of the #token and #juice regression coefficients in each area. Surprisingly, neuronal
196 populations in the AMY and VS showed a higher similarity in encoding #token and #juice (Figures 4I,
197 S3C; Fisher's z-transformation, $p < 0.05$). This result shows that the symbolic and primary reinforcers were
198 more similarly encoded in the AMY and VS. Although OFC encoded both at the highest level, it did so
199 with different population responses and, therefore, can discriminate these reinforcers best among the areas.

200 Value updates in the ventral striatum and orbitofrontal cortex

201 We used a stochastic reward schedule such that tokens were only updated in 75% of the trials, and token
202 losses only occurred when tokens could be lost. The encoding of Δtoken (Figure 3), therefore, may also
203 include the encoding of RPE (Figure 5A). To address this, we further fit the behaviors with a reinforcement
204 learning model (Figure 1G) and classified each neuron into Δtoken or RPE categories using linear
205 regression models (Figures S4A-B). RPE was calculated with the RW model. Neurons encoding Δtoken
206 responded to the token outcome, independent of the learning, whereas neurons encoding RPE encoded the
207 difference between the token outcome and the predicted token outcome, driving the learning. Overall, more
208 neurons in the OFC and VS encoded Δtoken or RPE than in the other two areas (Figures S4C-F; chi-square
209 test, $p < 0.001$). We then split neurons based on whether their regression coefficients were positive or
210 negative (e.g., whether they increased or decreased their firing rate for RPE; Figures 5B-E). For example,
211 a neuron was called a +RPE neuron when classified in the RPE category and with a positive regression
212 coefficient.

213 Although a similar proportion of neurons in the OFC and VS encoded these two types of value information,
214 they showed different patterns. Most neurons in the OFC at the time of token outcome encoded -Δtoken
215 and +RPE (Figure 5B). In other words, they fired more when losing more tokens (-Δtoken) or when they
216 did not lose tokens after choosing a loss option (+RPE). Specifically, this was aligned with the value
217 gradient carried by the choices (Figure 5B inset), even though it was a signal related to the outcome. On
218 the other hand, most neurons in the VS encoded +Δtoken and +RPE (Figure 5C). Neurons in the VS,
219 therefore, fired more for all good outcomes when gaining or not losing more tokens. This was well aligned
220 with the value gradient carried by the outcomes but not the choices (Figure 5C inset). This suggests that
221 value updates were referenced to outcomes in the VS and to choices in the OFC.

Flow of token information in the ventral network

Brain areas function as part of big networks but not as individual isolated areas. The task-relevant information also is not isolated in each area. To understand the flow of token information within the network, we measured the linear dynamics of the trajectories within a trial. This analysis used only simultaneously recorded ensembles and was carried out trial-by-trial. First, we projected the population neural activity from single trials into a 3-D subspace to generate the population state-space representation of #token, Δtoken, and cDir (Figure S2D). Then, we put these task-variable-specific latent variables together into a matrix. The columns included all the latent variables from all the areas, and the rows included each trial (Figure 6A). We stacked all the sessions (N = 16) with more than ten neurons recorded from each area simultaneously, then estimated the loading matrix A (Figure 6B). The matrix A characterized the flow of information among variables and areas (Figure S2E), with its columns representing the source areas and variables, and the rows representing the target areas and variables (Figure 6B). The matrices were estimated separately at each point in time, as a local linear approximation to the dynamics, which were likely nonlinear.

The eigenvalues of the A matrices captured the time constant of the dynamics of information flow among the areas in the network. The top two eigenvalues, which show two peaks, indicate robust information flow during the cue and token update epochs, and the third eigenvalue captured information flow during the token update epoch (Figure 6D). The left and right eigenvectors of the matrix A capture the temporal dynamics of output (Figures S5A-C) and input (Figures S5D-F) information. The first eigenvector also reflected the temporal structure of the task, showing phasic activity that was locked to the associated task variables. The values in the A matrices indicate the strength of information flow between a specific pair of areas and task variables. The temporal dynamics of the information flow varied across the time course of the trial (Figure 6C). We examined the flow of token information across areas and task variables (Figure S6) and found the strongest information flow existed within the same area for most conditions. This is consistent with other work showing that within-region dynamics are higher dimensional than across-region dynamics [26].

Our dynamics analysis showed good specificity between task variables. For example, despite the fact that cDir, #token, and Δtoken were represented across areas, there were minimal interactions between cDir and the value signals (Figure S6F-H). Aiming to address the flow of token information, we focused on the interactions within and between the Δtoken and #token. For Δtoken (Figure 6E), information flows showed peaks phase-locked to the token update. There was strong reciprocal information flow between the OFC

and MDt, reflecting the underlying anatomy [2]. We also found flow of information from AMY to the other areas, especially VS. For the #token (Figure 6F), the information flow was relatively continuous along the time course of the trial, which is consistent with the tokens being present on the screen across trials. More interestingly, there was a temporal derivative in the flow of information from #token to Δtoken in the OFC and VS (Figure 6G, 100 ms before vs. after derivative; Fisher's z-transformation; VS to VS, $p < 0.001$; OFC to OFC, $p < 0.05$). This shows the mechanism by which Δtoken was calculated in the network. Specifically, OFC and VS computed the difference in #token following and prior to the choice (i.e., $\Delta token = \#token_{after} - \#token_{before}$). Also, we found a stronger flow of information from the #token to Δtoken than from the Δtoken to #token (Figures S6B, D; Fisher's z-transformation, $p < 0.05$). This suggests that the updating of values in this four-area network originated predominantly in the VS and OFC, consistent with the single-cell results (Figures 5B-C). Because both areas showed the time-derivative dynamics, it seems likely that they may both be calculating the update, perhaps in parallel. Otherwise, one might expect a time-derivative in only one structure, paired with Δtoken flow between structures. During this calculation, phasic reciprocal information flow also existed between OFC-MDt and MDt-VS (Figure 6G), indicating that MDt bridged the information flow between the OFC and VS.

DISCUSSION

We examined the single-cell, population, and network representation of token-based RL across the ventral cortico-striatal-thalamo-cortical network [2,19,27]. We found that although a similar fraction of the population across these areas encoded the task variables, they encoded the information in different ways. VS showed a strong bias towards encoding appetitive choices and outcomes, including positive RPEs, with increased firing rates. On the other hand, OFC showed a more balanced encoding of positive and negative choices and outcomes with increased and decreased firing rates. OFC also strongly encoded primary and symbolic reinforcers but distinguished them at the population level. AMY, on the other hand, encoded both reinforcers similarly at the population level. AMY also showed enhanced salience coding and did not show correlations between choice and outcome value coding. Finally, MDt showed enhanced coding of actions and mediated information flow between the OFC and VS during token update calculations.

Primary and symbolic reinforcers

Human studies typically use symbolic reinforcers, while animal studies use primary reinforcers. However, similarities and differences between the population coding of primary and symbolic reinforcers have not been examined. AMY and, to some extent, VS populations showed stronger correlations between primary and symbolic reinforcers and, therefore, encoded tokens and primary reinforcers similarly. This may explain why earlier lesion work found AMY linked conditioned stimuli to the specific reward properties of the unconditioned stimuli they predicted [28]. OFC showed the most substantial coding of both primary and symbolic reinforcers but encoded them with the lowest similarity, therefore discriminating them well. Thus, OFC represents value information with high fidelity, using a code that preserves detailed information about the type and valence of reinforcement [29].

Gains and losses

We found that VS showed a strong bias towards monotonically encoding gains, specifically coding rewarding choices and outcomes, and better than the expected outcomes, with increased firing rates. Thus, VS consistently signaled token gains with increased firing rates. This is consistent with human imaging studies [30] and the longstanding suggestion that VS is important for motivation [31,32]. AMY neurons showed enhanced tuning for salience. The AMY population, uniquely among the areas, did not show a correlation between linear choice value and outcome value encoding, and showed negative correlations, at the population level, between gain and loss encoding, both of which are also consistent with salience coding. Previous studies found AMY responses to appetitive and aversive stimuli but could not assess salience

because they only used single outcomes for each valence [33,34]. OFC showed monotonic encoding of value across both positive and negative outcomes. Unlike VS and AMY, OFC neurons encoded values with both positive and negative slopes with a slight bias towards neurons responding more for negative value choices and loss outcomes. Recordings in the pregenual anterior cingulate cortex (ACC) [35] and the insula [21] similarly found neuronal populations coding appetitive and aversive choices with both positive and negative slopes. Thus, cortical areas show a more balanced coding of gains and losses than the VS and AMY.

OFC contributes to the encoding of symbolic reinforcers

We found that OFC encoded accumulated tokens and changes in tokens at both the single-cell and population levels. These findings are consistent with previous work showing that OFC codes state information about the environment [36-39], as tokens and token updates define states and state transitions in this task [40]. State representation is also referred to as a cognitive map, particularly when states have to be inferred [41,42]. It is a map because states are nodes on graphs, and one has to know not only the current state, but also subsequent states to which one can transition. Within reinforcement learning, states are the environmental variables relevant to the learning and action selection process [43]. We also found that OFC encoded choice value, which is consistent with previous work showing that OFC codes economic value [44]. However, we have found that choice value was broadly encoded across our network, and OFC did not encode value more strongly than other areas. We did find differences in the way in which OFC encoded values relative to the other areas, as discussed above. OFC has a high-fidelity representation of the state that includes negative outcomes and the strongest encoding of accumulated tokens.

The calculation of value updates

We also examined network computations within targeted information dimensions. Analysis of single-trial communication using neurons recorded from multiple regions simultaneously can provide evidence of dynamic network processes that underlie behaviors [45]. Most studies that have examined interactions between areas with neurophysiology data have focused on pairwise interactions between cortical regions and have also not identified dynamic signatures of computations. Instead, they have shown that interactions occur within specific dimensions and often during particular periods within a task. For example, an early study found that choice-relevant signals were relayed from the prefrontal to the parietal cortex to guide behavior [46]. A more recent study found that stronger value coding in the OFC led to accelerated ramping of signals in the ACC [47]. And we recently identified object-to-direction information flow among lateral prefrontal cortex subregions in a task in which the location of a valuable stimulus had to be identified before a saccade could be directed toward it to make a choice [48].

329 The current study measured the flow of value information using simultaneous population recordings from
330 four areas in the ventral network. This allowed us to control for multiple, although not all, inputs to each
331 area. We found that token update information was calculated in the OFC and VS as a time-derivative of the
332 information about accumulated tokens (i.e., $\Delta token_{t+1} = \#token_{t+1} - \#token_t$). The signal was earlier
333 in the OFC but stronger in the VS. Thus, the change in accumulated tokens drove information about token
334 outcomes. This is consistent with the finding that OFC and VS strongly encoded Δtoken and token
335 prediction errors at the single-cell level. Note that in previous work with probabilistic reward outcomes, we
336 did not find strong encoding of RPE in these structures [12,41]. We found minimal support for the alternative
337 hypothesis that token outcomes were integrated to generate information about accumulated tokens (i.e.,
338 $\#token_{t+1} = \#token_t + \Delta token_t$). Thus, analysis of neural dynamics identified an explicit computational
339 correlate of a behavioral process. We also found that MDt mediated interactions between the OFC and VS
340 with respect to token updates, by mediating reciprocal information flow between OFC-MDt and MDt-VS.
341 Interestingly, we did not identify a direct interaction between the OFC and the VS, even though they are
342 connected directly [49,50]. This suggests that the token updates were mediated within the cortical-thalamic-
343 striatal circuit but not the corticostriatal circuit.

344 Effects of loss of specific nodes on reinforcement learning

345 These results also provide insight into the task-dependent effects of lesions in previous work. For example,
346 tasks that use probabilistic delivery of primary reinforcers have shown deficits following lesions to all nodes
347 of the ventral network [5,6,10,11,13,51]. However, lesions of the AMY have almost no effect on learning in the
348 tokens task, and lesions of the VS only affect learning to discriminate gain magnitude but not gains vs.
349 losses [17,18]. Here, we found that AMY showed enhanced salience coding. Salience coding can be used as a
350 learning signal in the context of probabilistic reward outcomes because salience and gains relative to no
351 outcome are equivalent. However, salience coding cannot be used for learning in token-based reinforcement
352 because large gains and large losses are represented similarly. This may also explain why we found no
353 evidence for the effects of AMY lesions on unsigned prediction error variables in Pearce-Hall models of
354 RL that had been reported previously [52] when examined in probabilistic reward tasks [13]. We also found that
355 VS was strongly biased toward encoding appetitive outcomes. Therefore, lesions to the VS would be
356 expected to have a larger effect on discriminating among gains, as opposed to between gains and losses [17].
357 Although we did not record in the insula, recent work has shown that the insula has substantial coding of
358 losses in token-based decision-making [21], consistent with early fMRI results [53]. Therefore, it is possible that
359 manipulations of the insula would show effects specific to loss outcomes. Finally, we found that OFC may

preferentially mediate the calculation of token reinforcement, and therefore, manipulations of OFC may have large effects on learning in the tokens task.

Conclusion

Information about task variables was represented across the ventral network. Although all areas represented task variables, they did so differently. AMY encoded choice and outcome salience, and also encoded the primary and symbolic reinforcers similarly. VS and OFC encoded value information, with VS strongly biased towards coding positive outcomes with increased firing rates and OFC more balanced towards coding positive and negative outcomes. OFC and VS calculated Δtoken as the time-derivative of accumulated tokens, with a shorter latency in the OFC. MDt mediated the interactions about Δtoken between the OFC and VS during this process. Importantly, the computations relevant to symbolic reinforcement appear to be more strongly mediated by cortical structures, in this case, OFC, than subcortical structures, which may differ from primary reinforcement.

ACKNOWLEDGMENTS

We thank Christos Constantinidis, Vincent Costa and Diana Burk for their valuable comments, as well as Andy Mitz, Craig Taswell, Miriam Janssen and Sarah Falkovic for their technical help. This work was supported by the Intramural Research Program of the National Institute of Mental Health (ZIA MH002928) to B. B. A. and BBRF NARSAD Young Investigator Award (30892) to H. T. Anatomical MRI scanning was carried out in the Neurophysiology Imaging Facility Core (NIMH, NINDS, NEI).

AUTHOR CONTRIBUTIONS

Conceptualization, R. B., H. T., and B. B. A.; Methodology, H. T., R. B., and B. B. A.; Investigation, H. T., and R. B.; Visualization, H. T., and B. B. A.; Writing – Original Draft, H. T., and B. B. A.; Writing – Review & Editing, H. T. and B. B. A.; Funding Acquisition, H. T. and B. B. A.; Supervision, B. B. A.

DECLARATION OF INTERESTS

The authors declare no competing interests.

FIGURE LEGENDS

Figure 1. Task, behavior, recording method, and recorded areas.

(A-B) Behavioral task. (A) Structure of an individual trial. Successive frames illustrate the sequence of events. In each trial, only two of the four images were presented. Monkeys chose between them and gained or lost (stochastic: 75% change, 25% no change) the corresponding number of tokens. Choices could be made as soon as the images were shown. Accumulated tokens were cashed out every four to six trials, with one drop of juice for each token. (B) In each block, the monkeys learned the values (+2, +1, −1, −2) of four novel images.

(C) Schematic recording areas, including the orbitofrontal cortex (OFC), ventral striatum (VS), basolateral amygdala (AMY), and mediodorsal thalamus (MDt). Red arrows indicate the anatomical connections between these areas. The dashed line indicates an indirect connection.

(D-F) Semi-chronic recording. (D) A chamber was implanted. (E) Then, guide tubes were inserted into each target area through a grid. The guide tubes stopped above the target areas and stayed in the brain across recording days. The empty space in the chamber was filled with silicone sealant to prevent potential infections. Multi-site linear probes were lowered into target areas through the guide tubes every recording day. (F) A coronal section view of the chamber shows the shadow of a guide tube and the target area under MRI. The dashed green polygon indicates the silicone sealant.

(G-I) Choice behavior. (G) Fraction of choosing the image with a higher value in each condition. Blue lines represent monkey behavior, and green lines represent the prediction of the reinforcement learning model. Error bars represent mean ± SEM. The black asterisks indicate a significant difference in monkey behavior from the change level (one-sample t-test, $p < 0.01$). (H) The effect size of each task variable. Paired t-test, $^{**}p < 0.01$, $^{***}p < 0.001$. Non-significant pairs are not indicated. (I) Choice behavior as a function of the task variables. Results were averaged from two monkeys, n = 50 sessions. Violin plots showing the range of values calculated across all sessions (center dot, mean; box, 25th to 75th percentiles; whiskers, ± 1.5 × the interquartile range; dots, each session; shade, density curve). See also Figure S1.

410 ## Figure 2. Neural encoding of choices, primary and symbolic reinforcers.

411 **(A-F)** Percentage of neurons in each area encoding #token (A), Δtoken (B), #juice (C), the identity (D), *a*
412 *priori* value (E), and direction (F) of chosen images. Data were aligned to either the token update or the
413 onset of juice delivery. Inset in (B): ANOVA with 25 ms bin showing response latencies to Δtoken different
414 among areas. Dashed horizontal lines represent the chance level. Thick lines indicate a significant
415 difference between the corresponding area and chance level (binomial test, $p < 0.01$).

416 **(G-I)** Example neurons encoded the *a priori* value of the chosen images (G), #token (H), and #juice (I).
417 Each row in the raster plot represents the spikes during a trial. Red, green, blue, cyan, and magenta lines
418 represent fixation, cue onset, choice, token update, and juice delivery. The bars on the right side represent
419 the trial groups under different conditions. The bottom curves indicate mean activities, split by trial groups
420 defined by the bars.

421

422 ## Figure 3. Diverse representation of gains and losses.

423 **(A-E)** Percentage of neurons in each area encoding Δtoken. The insets indicate the corresponding tuning
424 curves. The coefficients could be positive or negative, meaning the tuning curves could be flipped vertically.
425 (A) General value: monotonic encoding values across gain and loss contexts. (B) Salience: monotonic
426 encoding values across gain and loss contexts but with opposite directions. (C) Gain/Loss: categorical
427 encoding values in gain and loss contexts. (D) Gain: encoding values only in the gain context. (E) Loss:
428 encoding values only in the loss context. Dashed horizontal lines represent chance levels, defined by the
429 mean of all areas during the fixation epoch. Thick lines indicate a significant difference between the
430 corresponding area and chance level (binomial test, $p < 0.05$). The double-colored asterisks indicate a
431 significant difference between pairs of areas indicated by the colors (chi-square test, $p < 0.05$).

432 **(F-J)** Example neurons for each category.

433 **(K-O)** Co-encoding of gains and losses. Correlations of the Gain and Loss regression coefficients at the
434 same (K) or cross (L-O) time points for each area. See also Figures S2 and S3.

435

436 ## Figure 4. Population encoding of value update and reinforcers.

437 **(A-G)** Co-encoding of cValue and Δtoken. (A-D) The x- and y-axes represent the regression coefficients
438 of cValue and Δtoken. The blue dots represent the neurons encoding both of them. The gray dots represent
439 the other neurons. Percentages on the left corner indicate the proportions of neurons encoding the cValue
440 and Δtoken. Black and red lines are the lines of best fit for all the neurons and significant neurons, and r
441 indicates the corresponding correlation coefficient. The bars summarize the distributions of blue and gray
442 dots, with blue and black lines representing the means. (E) Summary of the distributions of blue dots in
443 each quadrant (indicated in panel A) for each area. Chi-square test, *p < 0.05, **p < 0.01, ***p < 0.001.
444 (F) The proportion of neurons encoding cValue and Δtoken in each area. Summary of the percentages in
445 panels A-D. Chi-square test, **p < 0.01, ***p < 0.001. (G) Encoding similarity between cValue and Δtoken
446 in each area. Summary of the black r in panels A-D. Fisher's z-transformation, ***p < 0.001. Non-
447 significant pairs are not indicated.

448 (H-I) Co-encoding of #token and #juice. (H) The proportion of neurons encoding #token and #juice in each
449 area. Chi-square test, *p < 0.05, ***p < 0.001. (I) Encoding similarity between #token and #juice in each
450 area. Fisher's z-transformation, *p < 0.05, **p < 0.01, ***p < 0.001. See also Figure S3.

451

452 ## Figure 5. Value updates in the ventral striatum and orbitofrontal cortex.

453 **(A)** Relationship of *a priori* value, Δtoken, and RPE. v_h indicates the RW model-estimated value of the
454 image with *a priori* value h, which got closer to h as learning progressed.

455 **(B-E)** Percentage of neurons in each area encoding +Δtoken, -Δtoken, +RPE, and -RPE. Dashed horizontal
456 lines represent chance levels. Thick lines indicate a significant difference between the corresponding area
457 and chance level (binomial test, p < 0.025). The black asterisks indicate a significant difference among the
458 four categories (chi-square test, p < 0.01). Insets: OFC neurons encoded Δtoken and RPE in a manner
459 relevant to the value of choices. VS neurons encoded Δtoken and RPE correlate with the value of outcomes.
460 Thick lines represent more neurons. See also Figure S4.

461

Figure 6. Flow of token information in the ventral network.

(A-D) Detecting network communications with low dimensional information dynamics. (A) The task-related population response matrix consisted of single-trial-level population projections. Linear dynamic regression measured time-dependent flows of information across task variables and brain areas. (B) The loading matrix A maps the relationship between rows in matrix X. The x-axis represents the source areas and variables, and the y-axis represents the target areas and variables. For example, the values indicated by the red square represent the flow of Δtoken information from the MDt to #token, Δtoken, and cDir information in the AMY. (C) The continuous values in A matrices represent the strength of information flow across task variables and brain areas along the time course of the trial. (D) The eigenvalues of A matrices capture the strength of the overall information flow among the areas in the network.

(E-G) The flow of token information across areas, including within the Δtoken (E), #token (F), and between them (G). The titles indicate the source and target task variables. Each panel indicates one target area, and lines indicate source areas. Thick lines indicate a significant difference between the corresponding area and shuffled data (two sides z-test, $p < 0.01$). Shuffling was done by keeping the same conditions but shuffling the trial order of matrix X. See also Figures S2, S5, and S6.

STAR METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Neural data	This paper	DOI
Behavioral data	This paper	DOI
Software and algorithms		
MATLAB R2020b	MathWorks Inc.	https://www.mathworks.com/products/matlab.html
Analysis code	This paper	DOI
MonkeyLogic	NIMH/NIH	https://monkeylogic.nimh.nih.gov/index.html
Adobe Illustrator 2024	Adobe	https://www.adobe.com/products/illustrator.html
Wave_clus 3	Chaure et al., 2018[54]	https://github.com/csn-le/wave_clus
Other		
V-probe	Plexon	https://plexon.com/product-category/v-probes/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Bruno B. Averbeck (bruno.averbeck@nih.gov).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The datasets supporting the current study will be publicly available as of the date of publication. DOIs are listed in the key resources table.

All original codes will be publicly available as of the date of publication. DOIs are listed in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

492 ## EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

493 ### Subjects

494 The experiments were performed on one adult male (10 kg) and one adult female (7.5 kg) rhesus macaque
495 (*Macca mulatta*). They were 8-10 years old. The monkeys were pair-housed when possible and had access
496 to food 24 hours per day. On testing days, the monkeys were placed on water control and earned their juice
497 through performing the task. On non-testing days, the monkeys were given *ad libitum* access to water.

498 ## METHOD DETAILS

499 ### Ethics

500 Experimental procedures for all monkeys followed *the Guide for the Care and Use of Laboratory Animals*
501 and were approved by the National Institute of Mental Health Animal Care and Use Committee.

502 ### Experimental Setup

503 Monkeys were trained to perform a saccade-based two-armed bandit task. Stimuli were presented on a 19-
504 inch LCD monitor situated 40 cm from the monkeys' eyes. During training and testing, the monkeys sat in
505 a primate chair with their heads restrained. Stimulus presentation and behavioral monitoring were
506 controlled by MonkeyLogic [55]. The eye movements were monitored at 400 fps using a Viewpoint eye
507 tracker (Arrington Research, Scottsdale, AZ) and sampled at 1 kHz. A fixed amount of apple juice was
508 delivered through a pressurized plastic tube gated by a solenoid valve on rewarded trials.

509 ### Task Design

510 The task was developed and first used in our previous study [17]. Each session had nine blocks. Each block
511 used four novel images associated with different values, including +2, +1, −1, and −2. The monkeys
512 obtained more tokens by choosing images associated with positive values. Choosing one of the images led
513 to gaining or losing a corresponding number of tokens. Also, token numbers could not be negative, so
514 choosing a loss image when there were no accumulated tokens had no effect. We used a stochastic reward
515 schedule. The number of tokens updated 75% of the time and did not change 25% of the time. To complete
516 a trial successfully, the monkey first acquired and held central fixation for 400-600 ms. Then, two of the
517 four images were randomly selected by the computer and displayed on the screen. The animal made its
518 selection by saccading to one of them. The unchosen image disappeared when the monkey reached the

target. Saccade fixation was maintained on the chosen image for 500 ms. After that, the chosen image also disappeared, and the token number was updated according to the chosen image. Tokens accumulated across trials and were cashed out for juice every four to six trials, with the interval randomly selected. At cash-out, the animals were given one drop of juice for each token. When each drop of juice was delivered, one token was removed from the screen. Up to 12 tokens were accumulated and were displayed on the screen across trials.

The task had six individual conditions defined by the possible values of the image pairs. The conditions within a block of 108 trials (6 conditions × 2 counterbalanced for left and right sides × 9 repetitions) were presented pseudo-randomly. The animals saw each condition twice, once on the left and once on the right, every 12 trials before seeing any condition a third time. We introduced four novel images at the beginning of each block. Images provided as choice options were normalized for luminance and spatial frequency using the SHINE toolbox for MATLAB, described previously [17].

Surgical procedures

Each monkey was surgically implanted with a titanium headpost, and a 25 × 35 mm recording chamber to allow vertical grid access to the OFC, VS, AMY, and MDt (Figures 1D-F). Grid holes for the MDt had 16° angles to allow better access to the target area (Figure S1D). Chamber placements were planned and verified with T1 and T2 magnetic resonance imaging (MRI, 3.0 T). Small burr holes were drilled above each target area. A grid was installed, and one guide tube was inserted through each burr hole. The guide tubes were lowered to about 1-4 mm (depending on the target areas) above the target areas and were glued to the grid. We removed the guide tubes and placed new ones in the adjacent locations after 3-5 recording days (Figures S1D-E). The locations of the guide tubes were verified with MRI after guide tube replacement. All sterile surgeries and MRI scans were performed under anesthesia.

Neurophysiological Recordings

Neurophysiology recordings (Figures S1C-D) began after the monkeys had recovered from the surgery. We lowered one linear electrode array (V-probe, Plexon Inc, Dallas, TX) into each guide tube on every recording day. Thirty-two channel electrodes with 150 μm inter-contact spacing probes were used in the VS and MDt, and 64-channel electrodes with 150 μm inter-contact spacing probes were used in the OFC and AMY. The probes were advanced to their target location by a four-channel micromanipulator (NAN Instruments, Nazareth, Israel) attached to the recording chamber. The depths of the neurons were estimated by their recording locations relative to the tip of the guide tubes (verified with MRI). Electrophysiological

data were acquired with a 512-channel Grapevine System (Ripple, Salt Lake City, UT). The spike acquisition threshold was set at a 4.0 × root mean square (RMS) of the baseline signal for each electrode. Behavioral event markers from MonkeyLogic and eye-tracking signals from Viewpoint were sent to the Ripple acquisition system. The extracellular signals were high-pass filtered (1 kHz cutoff) and digitized at 30 kHz to acquire the single-cell activity. Spikes were sorted offline via Wave_clus 3 [54].

Choice behavior

Each block had six conditions. The conditions pseudo-randomly appeared in the task. The number of observations in each condition increased as a function of learning. We quantified choice behavior during the task by measuring the fraction of choosing the image associated with a higher value in each condition. We then measured how different task variables affected the choice behavior by fitting a multi-way ANOVA model. Factors, including the number of observations (Trial), Gain/Loss condition (Loss/Loss, Loss/Gian, or Gain/Gain), value difference of the options (Δvalue, e.g., Δvalue = 4 in the condition of -2 vs. +2), the number of tokens before choice (#token), and the number of trials since last token cashout (CashID) were used in the model. The model was run session by session to measure the contribution of each variable in each session, using all the sessions to acquire the statistics for each variable. All trials in which monkeys chose one of the two stimuli were analyzed. Trials in which the monkey broke fixation, failed to make a choice, or attempted to saccade to more than one target were excluded.

Effect size

To quantify the contribution of each task variable to the monkey's choice. We computed each factor's effect size, ω^2, from the ANOVA model output. $\omega^2$is an unbiased estimator of the amount of variance in neural activity explained by each task variable, and ranges between -1 and 1 [56]. It is given by:

$$\omega^2 = \frac{df_{effect} \times (MS_{effect} - MS_{error})}{SS_{total} + MS_{error}}$$

where df_{effect} refers to the degrees of freedom associated with the factor, MS_{effect} refers to the mean squares, MS_{error} is the mean squared error, SS_{total} is the sum of squares of all factors.

Responsive neurons

To identify neural responses to different task components, we fit a sliding window multi-way ANOVA model to spike counts computed in 200 ms bins, advanced in 50 ms increments, and time-locked to token

update. Factors, including the number of tokens on the screen (#token, may change after the choices), the change of token numbers (Δtoken), the drops of juice delivered (#juice), image pairs with different value combinations presented (condition, e.g., +2 vs. -1), the order of blocks (blockID), the identity (cStim), *a priori* value (cValue), and direction (cDir) of chosen images were used in the model. The stimulus identity was the specific image used in each block to represent each outcome, which was the interaction of cValue and blockID. The factor blockID was used to remove non-stationarity due to drift.

Statistic test for the proportion of neurons

The binomial test was applied to test whether the proportion of responsive neurons was significantly above the chance level (5% most of the time). The chi-square test was used to compare the proportions of responsive neurons between different pairs of brain areas or among four areas. Significant encoding for each factor at each time bin was evaluated at $p < 0.05$. A neuron that showed a significant response to a factor in no less than three contiguous bins in the statistics was considered to be responsive to that factor.

Neuronal coding regression analysis

To quantitatively characterize how the updating of values (including cValue and Δtoken) was encoded in different areas, we examined the activity of each neuron using a series of multivariate linear regression models [21]. The dependent variable, neural activity, was first z-scored by subtracting the mean response from the firing rate at each time and in each trial and dividing the result by the standard deviation of the responses. Both the mean and the standard deviation were computed by combining the neurons' responses across all trials and times. We then described the z-scored responses of neuron i at time t as a linear combination of several task variables. The independent variables were the same as the factors used in the ANOVA model:

$$r_{i,t}(k) = \beta_{i,t,\mathrm{cValue}} * cValue(k) + \beta_{i,t,\Delta token} * \Delta token(k) + \cdots + \beta_{i,t,cDir} * cDir(k) + \varepsilon$$

where $r_{i,t}(k)$ is the z-scored response of neuron i at time t and on trial k, $\Delta token(k)$ is the change of tokens on trial k. The regression coefficients, $\beta_{i,t,f}$, describe how much the trial-by-trial firing rate of neuron i, at a given time t during the trial, depends on the corresponding task variable f.

We tested all potential combinations of tuning for Δtoken (six forms: General value, Salience, Gain/Loss, Gain, Loss, nan) and cValue (six forms: General value, Salience, Gain/Loss, Gain, Loss, nan). Nan means the variable was removed from the model. A particular variable could only be represented by one specific form but not by combinations of more than one form (e.g., a model containing the Δtoken variable could

include Gain or Salience but not both). In total, we tested 6 × 6 = 36 models for each neuron. This method is also illustrated in Figure S2A. We determined the best-fitting model for each neuron using the Akaike information criterion (AIC). Neurons were classified into different functional categories (up to one category per variable) according to the combination of the forms of these two variables included in the best-fitting model. We computed the AIC:

$$AIC = 2p - 2\ln(\hat{L})$$

where p is the number of free parameters in the model, and $\hat{L}$ is the maximized value of the likelihood function.

Population coding similarity

The regression coefficient, $\beta_{i,\mathrm{t},f}$, from the multivariate linear regression model reflects the weight of one task variable, f, in explaining the variation of the neuron's activity at time t. The regression coefficients, $\overrightarrow{\beta_f}$, from a neural population, represent the weights of each neuron in the population encoding one task variable. Hence, computing the correlation between two regression coefficient vectors from the same neural population tells us how similar a neural population encodes two task variables (Figures S2B-C):

$$\mathrm{r} = \mathrm{corr}(\overrightarrow{\beta_{f1}}, \overrightarrow{\beta_{f2}})$$

where $\overrightarrow{\beta_f}$ is a vector that consists of regression coefficients of task variable f in a neural population. To assess the significance of the correlation between two regression coefficients, we transformed the correlation coefficients into normally distributed z-scores using Fisher's z-transformation.

Reward prediction error

The RPE was defined as the difference between the change of tokens $\Delta token(t)$ and the estimated value of the chosen image, which is given by:

$$RPE(t) = \Delta token(t) - v_i(t)$$

where t means trial order, and i means the chosen option among two options. The updating of value v_i was estimated using the Rescorla–Wagner equation:

$$v_i(t+1) = v_i(t) + \rho_h RPE(t)$$

where ρ is the learning rate. The updated value estimate $v_i(t+1)$ equals the previous value estimate $v_i(t)$ plus the RPE scaled by the cue-dependent learning rate ρ_h for images associated with different *a priori* values h. These values were then passed through a soft-max function to give choice probabilities for the image pairs presented in each trial:

$$d_j(t) = (1 + e^{\beta(v_i(t) - v_j(t))})^{-1}, d_i(t) = 1 - d_j(t)$$

where β is the choice consistency or inverse temperature parameter, fit across all six cue conditions, and i and j are the two choice options. We then maximized the likelihood of the animal's choices, D, given the parameters, using the cost function:

$$f(D|\alpha_i, \beta) = \prod_t [d_1(t)c_1(t) + d_2(t)c_2(t)$$

where $d_1(t)$ is the choice probability value for one option on trial t, and $c_1(t)$ and $c_2(t)$ are indicator variables that take on a value of 1 if the corresponding option was chosen, and 0 otherwise. This model was fit across blocks in each session for each monkey to give one set of free parameters for each session.

Targeted dimensionality reduction

We used the regression coefficients described above to identify dimensions in state space representing each task variable [57]. For each variable, we first build a set of coefficient vectors $\overrightarrow{\beta_{f,t}}$ whose entries $\beta_{f,t}(i)$ correspond to the regression coefficient for task variable f, time t, and neuron i. The vectors $\overrightarrow{\beta_{f,t}}$ (of length N_{unit}) are obtained by simply rearranging the entries of the vectors $\overrightarrow{\beta_{\iota,t}}$ (of length N_f) computed above. Each vector, $\overrightarrow{\beta_{f,t}}$, thus corresponds to a direction in state space that accounts for variance in the population response at time t, due to variation in task variable f.

We used time-dependent regression vectors, B_t, which is a matrix with each column corresponding to the one regression vector $\overrightarrow{\beta_{f,t}}$. We referred to them as the 'task-related axes'. These axes span a 'regression subspace' representing the task-related information the neural population coded. We then projected the single-trial population responses onto these axes to study the representation of the task-related variables in each trial:

$$X_t = {B_t}^T R_t$$

where R_t is single-trial population response at time t, which is of dimension $N_{unit} \times N_{trial}$. And X_t is the set of time-series vectors over all task variables and trials, which is of dimension $N_f \times N_{trial}$. It represents the population coding of task-related information at time t of every single trial. This method is also illustrated in Figure S2D.

Linear dynamics

We fit an autoregressive, linear model to the X_t,

$$X(t+1) = AX(t) + \varepsilon(t)$$

where $X(t) = [X_{OFC}(t), X_{VS}(t), X_{AMY}(t), X_{MDt}(t)]^T$ is a matrix of dimension $(N_{area} \cdot N_f) \times N_{trial}$, and $X_{area}(t) = [x_{k,\#token}(t), x_{k,\Delta token}(t), x_{k,cDir}(t)]$ is a matrix of dimension $N_{trial} \times N_f$. $\varepsilon(t)$ is the noise term. A is a 12 × 12 dynamics matrix, where the 12 = 4 × 3 dimensions correspond to the combinations of brain areas and task variables. The model was fit using data points in a non-overlapping 25 ms moving window. This resulted in a time-dependent estimate of the matrix A. Eigenvalues of A closer to 0 indicate a faster decay, and eigenvalues near 1 would correspond to a slower decay. The method is also illustrated in Figures S2E and 6A-C.

A z-test was applied to test for significant differences between the actual and shuffled results for the elements of the A matrices. The shuffled data were generated using the same matrix X but with randomized trial indexes across rows ($areas \times variables$). The actual results with a mean outside the 95% confidence interval of the shuffled results showed a significant response.

QUANTIFICATION AND STATISTICAL ANALYSIS

Unless otherwise indicated, all data were presented as means ± SEM (standard error of the mean). The statistical analyses performed were indicated in the main text and detailed in STAR Methods. Statistical comparisons were analyzed in MATLAB (Mathworks), including parametric, non-parametric, and permutation-based statistics, as detailed in STAR Methods. Figures were prepared with MATLAB.

REFERENCES

1. Alexander, G.E., DeLong, M.R., and Strick, P.L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annu. Rev. Neurosci. *9*, 357-381.
2. Averbeck, B.B., and Murray, E.A. (2020). Hypothalamic Interactions with Large-Scale Neural Circuits Underlying Reinforcement Learning and Motivated Behavior. Trends Neurosci. *43*, 681-694.
3. Kim, C.K., Adhikari, A., and Deisseroth, K. (2017). Integration of optogenetics with complementary methodologies in systems neuroscience. Nat. Rev. Neurosci. *18*, 222-235.
4. Rich, E.L., and Averbeck, B.B. (2024). Neurophysiology: Differentiating Functional Contributions across Prefrontal Cortex. In The Frontal Cortex: Organization, Networks, and Function, M.T. Banich, S.N. Haber, and T.W. Robbins, eds. (MIT Press).
5. Chakraborty, S., Kolling, N., Walton, M.E., and Mitchell, A.S. (2016). Critical role for the mediodorsal thalamus in permitting rapid reward-guided updating in stochastic reward environments. Elife *5*. 10.7554/eLife.13588.
6. Walton, M.E., Behrens, T.E., Buckley, M.J., Rudebeck, P.H., and Rushworth, M.F. (2010). Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. Neuron *65*, 927-939.
7. Betzel, R.F., and Bassett, D.S. (2017). Multi-scale brain networks. NeuroImage *160*, 73-83.
8. Chau, B.K., Sallet, J., Papageorgiou, G.K., Noonan, M.P., Bell, A.H., Walton, M.E., and Rushworth, M.F. (2015). Contrasting Roles for Orbitofrontal Cortex and Amygdala in Credit Assignment and Learning in Macaques. Neuron *87*, 1106-1118.
9. Houk, J.C., Adams, J.L., and Barto, A.G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In Models of information processing in the basal ganglia., (The MIT Press), pp. 249-270.
10. Rudebeck, P.H., Saunders, R.C., Lundgren, D.A., and Murray, E.A. (2017). Specialized Representations of Value in the Orbital and Ventrolateral Prefrontal Cortex: Desirability versus Availability of Outcomes. Neuron *95*, 1208-1220 e1205.
11. Rudebeck, P.H., Saunders, R.C., Prescott, A.T., Chau, L.S., and Murray, E.A. (2013). Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. Nat. Neurosci. *16*, 1140-1145.
12. Costa, V.D., Mitz, A.R., and Averbeck, B.B. (2019). Subcortical Substrates of Explore-Exploit Decisions in Primates. Neuron *103*, 533-545 e535.
13. Costa, V.D., Dal Monte, O., Lucas, D.R., Murray, E.A., and Averbeck, B.B. (2016). Amygdala and Ventral Striatum Make Distinct Contributions to Reinforcement Learning. Neuron *92*, 505-517.
14. Rothenhoefer, K.M., Costa, V.D., Bartolo, R., Vicario-Feliciano, R., Murray, E.A., and Averbeck, B.B. (2017). Effects of Ventral Striatum Lesions on Stimulus-Based versus Action-Based Reinforcement Learning. J. Neurosci. *37*, 6902-6914.
15. Joel, D., Niv, Y., and Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. Neural Netw. *15*, 535-547, Pii s0893-6080(02)00047-3.
16. Takahashi, Y., Schoenbaum, G., and Niv, Y. (2008). Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model. Front. Neurosci. *2*, 86-99.
17. Taswell, C.A., Costa, V.D., Murray, E.A., and Averbeck, B.B. (2018). Ventral striatum's role in learning from gains and losses. Proc Natl Acad Sci U S A *115*, E12398-E12406.
18. Taswell, C.A., Janssen, M., Murray, E.A., and Averbeck, B.B. (2023). The motivational role of the ventral striatum and amygdala in learning from gains and losses. Behav. Neurosci. *137*, 268-280.

19. Averbeck, B., and O'Doherty, J.P. (2022). Reinforcement-learning in fronto-striatal circuits. Neuropsychopharmacology *47*, 147-162.
20. Seo, H., and Lee, D. (2009). Behavioral and neural changes after gains and losses of conditioned reinforcers. J. Neurosci. *29*, 3627-3641.
21. Yang, Y.P., Li, X., and Stuphorn, V. (2022). Primate anterior insular cortex represents economic decision variables proposed by prospect theory. Nat Commun *13*, 717.
22. Tang, H., Costa, V.D., Bartolo, R., and Averbeck, B.B. (2022). Differential coding of goals and actions in ventral and dorsal corticostriatal circuits during goal-directed behavior. Cell Rep. *38*, 110198.
23. Parker, N.F., Baidya, A., Cox, J., Haetzel, L.M., Zhukovskaya, A., Murugan, M., Engelhard, B., Goldman, M.S., and Witten, I.B. (2022). Choice-selective sequences dominate in cortical relative to thalamic inputs to NAc to support reinforcement learning. Cell Rep. *39*, 110756.
24. Bartolo, R., and Averbeck, B.B. (2020). Prefrontal Cortex Predicts State Switches during Reversal Learning. Neuron *106*, 1044-1054 e1044.
25. Blanco-Pozo, M., Akam, T., and Walton, M.E. (2024). Dopamine-independent effect of rewards on choices through hidden-state inference. Nat. Neurosci. *27*, 286-297.
26. Semedo, J.D., Zandvakili, A., Machens, C.K., Yu, B.M., and Kohn, A. (2019). Cortical Areas Interact through a Communication Subspace. Neuron *102*, 249-259 e244.
27. Averbeck, B.B., and Costa, V.D. (2017). Motivational neural circuits underlying reinforcement learning. Nat. Neurosci. *20*, 505-512.
28. Cardinal, R.N., Parkinson, J.A., Hall, J., and Everitt, B.J. (2002). Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. Neurosci Biobehav Rev *26*, 321-352.
29. McDannald, M.A., Esber, G.R., Wegener, M.A., Wied, H.M., Liu, T.L., Stalnaker, T.A., Jones, J.L., Trageser, J., and Schoenbaum, G. (2014). Orbitofrontal neurons acquire responses to 'valueless' Pavlovian cues during unblocking. Elife *3*, e02653.
30. Yacubian, J., Glascher, J., Schroeder, K., Sommer, T., Braus, D.F., and Buchel, C. (2006). Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. J. Neurosci. *26*, 9530-9537.
31. Mogenson, G.J., Jones, D.L., and Yim, C.Y. (1980). From motivation to action: functional interface between the limbic system and the motor system. Prog. Neurobiol. *14*, 69-97.
32. Shiflett, M.W., and Balleine, B.W. (2010). At the limbic-motor interface: disconnection of basolateral amygdala from nucleus accumbens core and shell reveals dissociable components of incentive motivation. Eur. J. Neurosci. *32*, 1735-1743.
33. Belova, M.A., Paton, J.J., Morrison, S.E., and Salzman, C.D. (2007). Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala. Neuron *55*, 970-984.
34. Paton, J.J., Belova, M.A., Morrison, S.E., and Salzman, C.D. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. Nature *439*, 865-870.
35. Amemori, K., and Graybiel, A.M. (2012). Localized microstimulation of primate pregenual cingulate cortex induces negative decision-making. Nat. Neurosci. *15*, 776-785.
36. Schuck, N.W., Cai, M.B., Wilson, R.C., and Niv, Y. (2016). Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. Neuron *91*, 1402-1412.
37. Basu, R., Gebauer, R., Herfurth, T., Kolb, S., Golipour, Z., Tchumatchenko, T., and Ito, H.T. (2021). The orbitofrontal cortex maps future navigational goals. Nature *599*, 449-452.
38. Costa, K.M., Scholz, R., Lloyd, K., Moreno-Castilla, P., Gardner, M.P.H., Dayan, P., and Schoenbaum, G. (2023). The role of the lateral orbitofrontal cortex in creating cognitive maps. Nat. Neurosci. *26*, 107-115.
39. Gardner, M.P.H., and Schoenbaum, G. (2021). The orbitofrontal cartographer. Behav. Neurosci. *135*, 267-276.

40. Burk, D.C., Taswell, C., Tang, H., and Averbeck, B.B. (2023). Computational mechanisms underlying motivation to earn symbolic reinforcers. bioRxiv, 2023.2010. 2011.561900.
41. Costa, V.D., and Averbeck, B.B. (2020). Primate Orbitofrontal Cortex Codes Information Relevant for Managing Explore-Exploit Tradeoffs. J. Neurosci. *40*, 2553-2561.
42. Wilson, R.C., Takahashi, Y.K., Schoenbaum, G., and Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. Neuron *81*, 267-279.
43. Averbeck, B.B. (2015). Theory of choice in bandit, information sampling and foraging tasks. PLoS Comput. Biol. *11*, e1004164, e1004164.
44. Padoa-Schioppa, C., and Cai, X. (2011). The orbitofrontal cortex and the computation of subjective value: consolidated concepts and new perspectives. Ann. N. Y. Acad. Sci. *1239*, 130-137.
45. Vinck, M., Uran, C., Spyropoulos, G., Onorato, I., Broggini, A.C., Schneider, M., and Canales-Johnson, A. (2023). Principles of large-scale neural interactions. Neuron *111*, 987-1002.
46. Crowe, D.A., Goodwin, S.J., Blackman, R.K., Sakellaridi, S., Sponheim, S.R., MacDonald, A.W., 3rd, and Chafee, M.V. (2013). Prefrontal neurons transmit signals to parietal neurons that reflect executive control of cognition. Nat. Neurosci. *16*, 1484-1491.
47. Balewski, Z.Z., Elston, T.W., Knudsen, E.B., and Wallis, J.D. (2023). Value dynamics affect choice preparation during decision-making. Nat. Neurosci. *26*, 1575-1583.
48. Tang, H., Bartolo, R., and Averbeck, B.B. (2021). Reward-related choices determine information timing and flow across macaque lateral prefrontal cortex. Nat Commun *12*, 894.
49. Haber, S.N., and Knutson, B. (2010). The reward circuit: linking primate anatomy and human imaging. Neuropsychopharmacology *35*, 4-26.
50. Averbeck, B.B., Lehman, J., Jacobson, M., and Haber, S.N. (2014). Estimates of projection overlap and zones of convergence within frontal-striatal circuits. J. Neurosci. *34*, 9497-9505.
51. Rudebeck, P.H., Ripple, J.A., Mitz, A.R., Averbeck, B.B., and Murray, E.A. (2017). Amygdala Contributions to Stimulus-Reward Encoding in the Macaque Medial and Orbital Frontal Cortex during Learning. J. Neurosci. *37*, 2186-2202.
52. Li, J., Schiller, D., Schoenbaum, G., Phelps, E.A., and Daw, N.D. (2011). Differential roles of human striatum and amygdala in associative learning. Nat. Neurosci. *14*, 1250-1252.
53. Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., and Frith, C.D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature *442*, 1042-1045.
54. Chaure, F.J., Rey, H.G., and Quian Quiroga, R. (2018). A novel and fully automatic spike-sorting implementation with variable number of features. J. Neurophysiol. *120*, 1859-1871.
55. Hwang, J., Mitz, A.R., and Murray, E.A. (2019). NIMH MonkeyLogic: Behavioral control and data acquisition in MATLAB. J. Neurosci. Methods *323*, 13-21.
56. Olejnik, S., and Algina, J. (2000). Measures of Effect Size for Comparative Studies: Applications, Interpretations, and Limitations. Contemp. Educ. Psychol. *25*, 241-286.
57. Mante, V., Sussillo, D., Shenoy, K.V., and Newsome, W.T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature *503*, 78-84.

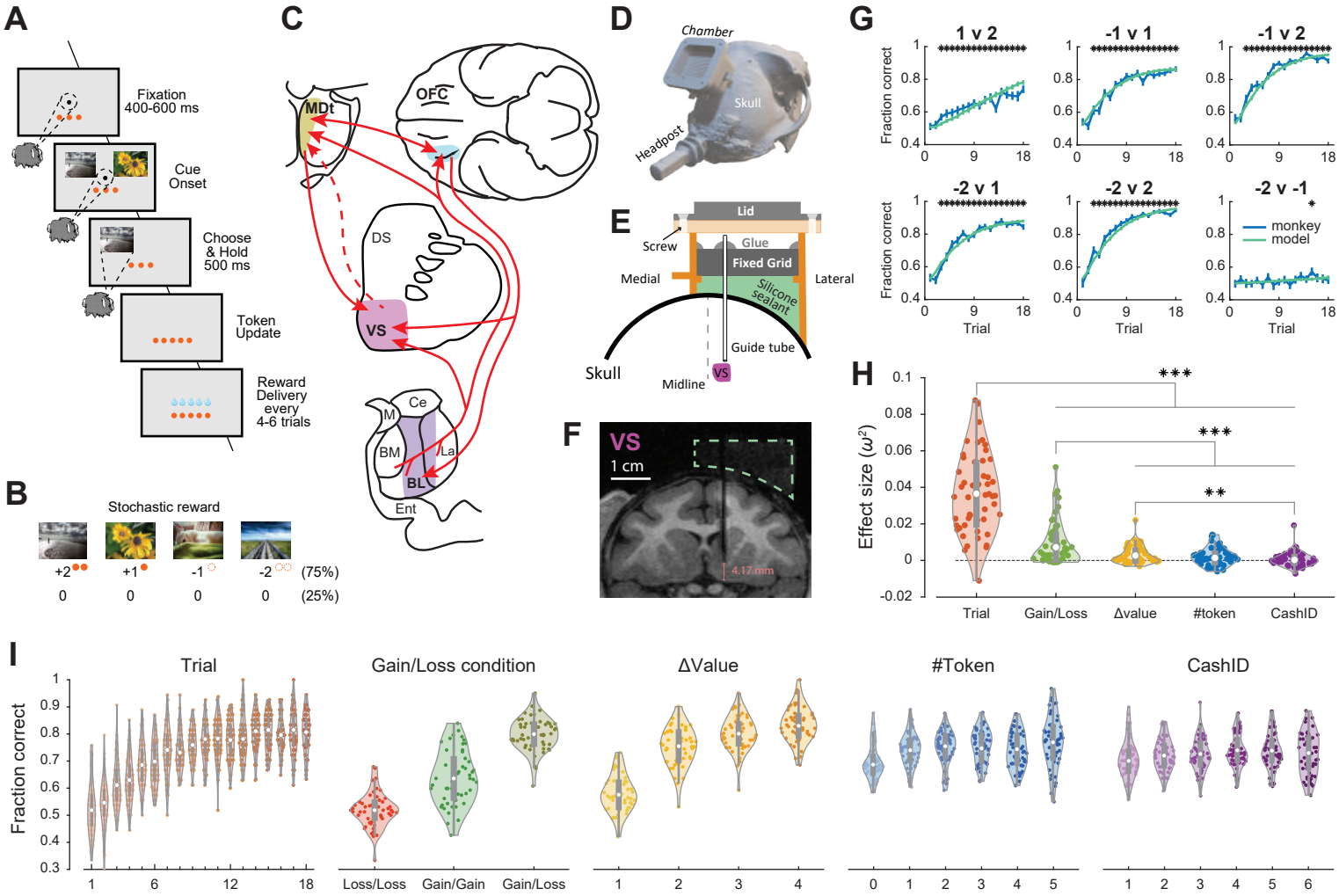
A
Fixation
400-600 ms
Cue
Onset
Choose
& Hold
500 ms
Token
Update
Reward
Delivery
every
4-6 trials
B
Stochastic reward
+2
+1
-1
-2
(75%)
0
0
0
0
(25%)
C
MDt
OFC
DS
VS
Ce
M
BM
La
BL
Ent
D
Chamber
Skull
Headpost
E
Lid
Screw
Glue
Fixed Grid
Medial
Lateral
Silicone sealant
Guide tube
Skull
Midline
VS
F
VS
1 cm
4.17 mm
G
1 v 2
-1 v 1
-1 v 2
-2 v 1
-2 v 2
-2 v -1
Fraction correct
Trial
monkey
model
H
Effect size (ω^2)
Trial
Gain/Loss
Δvalue
#token
CashID
I
Trial
Gain/Loss condition
ΔValue
#Token
CashID
Fraction correct
Loss/Loss
Gain/Gain
Gain/Loss

Figure 1

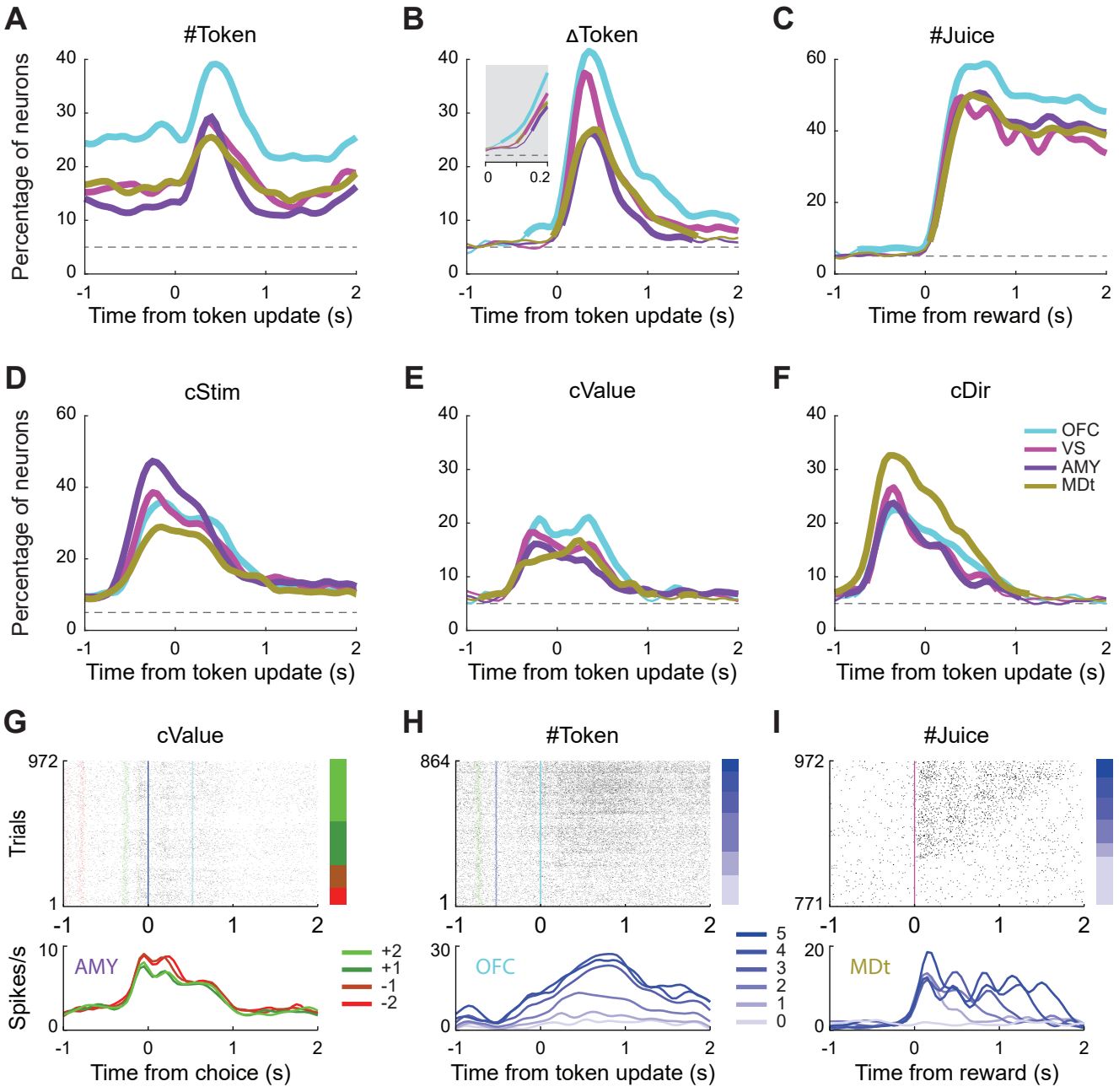
A
#Token
Percentage of neurons
Time from token update (s)
B
ΔToken
Time from token update (s)
C
#Juice
Time from reward (s)
D
cStim
Percentage of neurons
Time from token update (s)
E
cValue
Time from token update (s)
F
cDir
OFC
VS
AMY
MDt
Time from token update (s)
G
cValue
Trials
Spikes/s
AMY
+2
+1
-1
-2
Time from choice (s)
H
#Token
OFC
5
4
3
2
1
0
Time from token update (s)
I
#Juice
MDt
Time from reward (s)

Figure 2

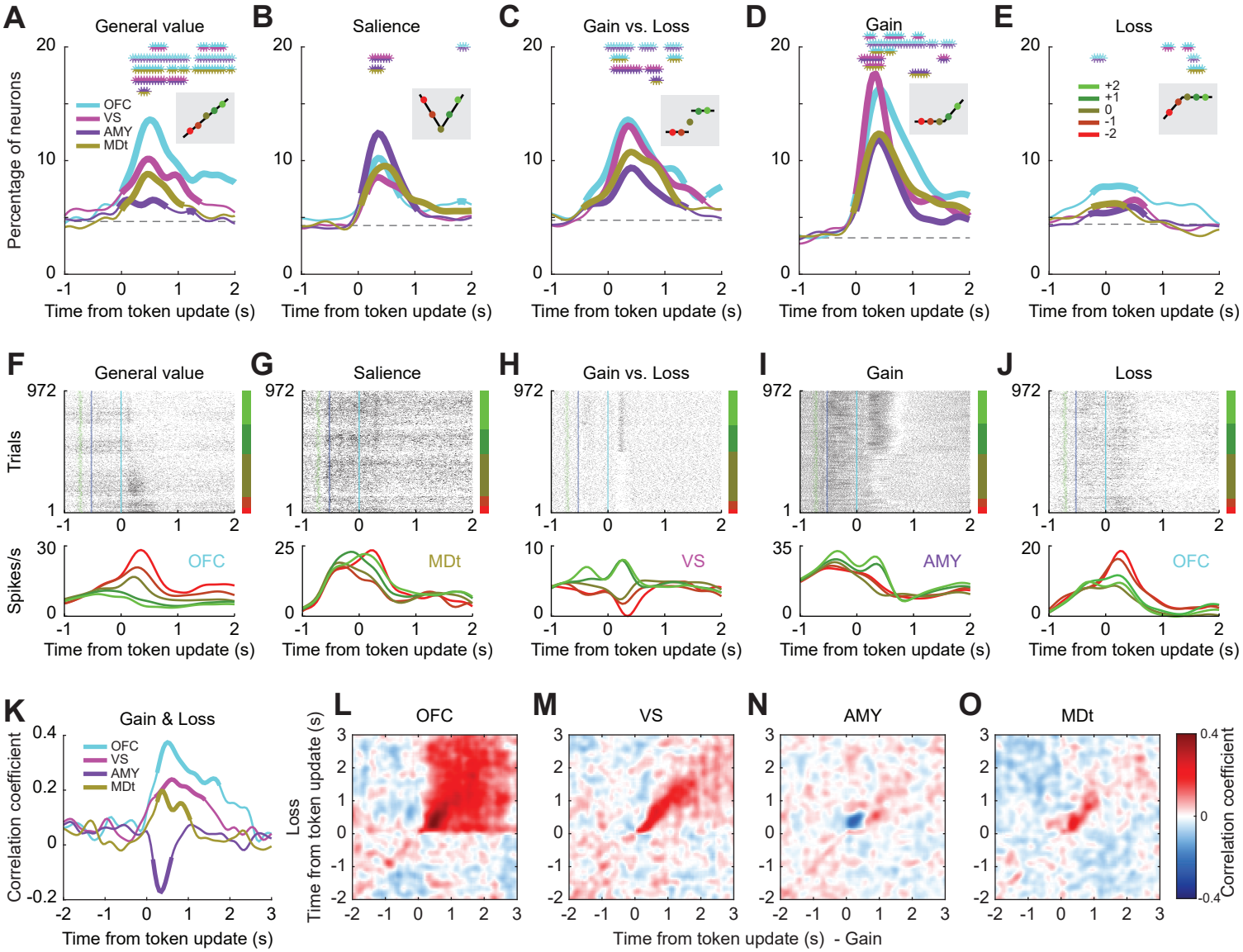
A
General value
B
Salience
C
Gain vs. Loss
D
Gain
E
Loss
Percentage of neurons
OFC
VS
AMY
MDt
+2
+1
0
-1
-2
Time from token update (s)
F
G
H
I
J
Trials
Spikes/s
K
Gain & Loss
Correlation coefficient
L
M
N
O
Time from token update (s) - Gain
Correlation coefficient

Figure 3

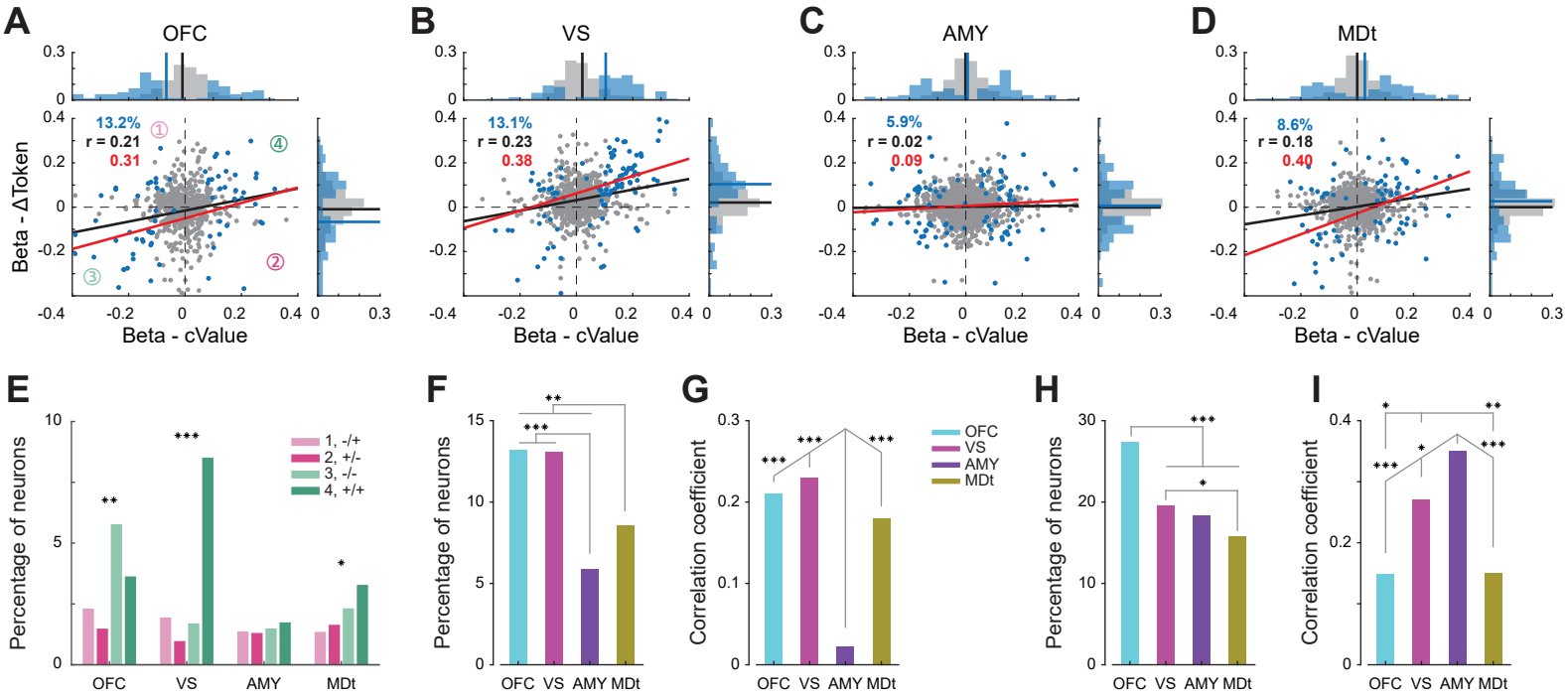
A
OFC
13.2%
r = 0.21
0.31
Beta - ΔToken
Beta - cValue
B
VS
13.1%
r = 0.23
0.38
C
AMY
5.9%
r = 0.02
0.09
D
MDt
8.6%
r = 0.18
0.40
E
Percentage of neurons
1, -/+
2, +/-
3, -/-
4, +/+
OFC
VS
AMY
MDt
F
G
Correlation coefficient
H
I

Figure 4

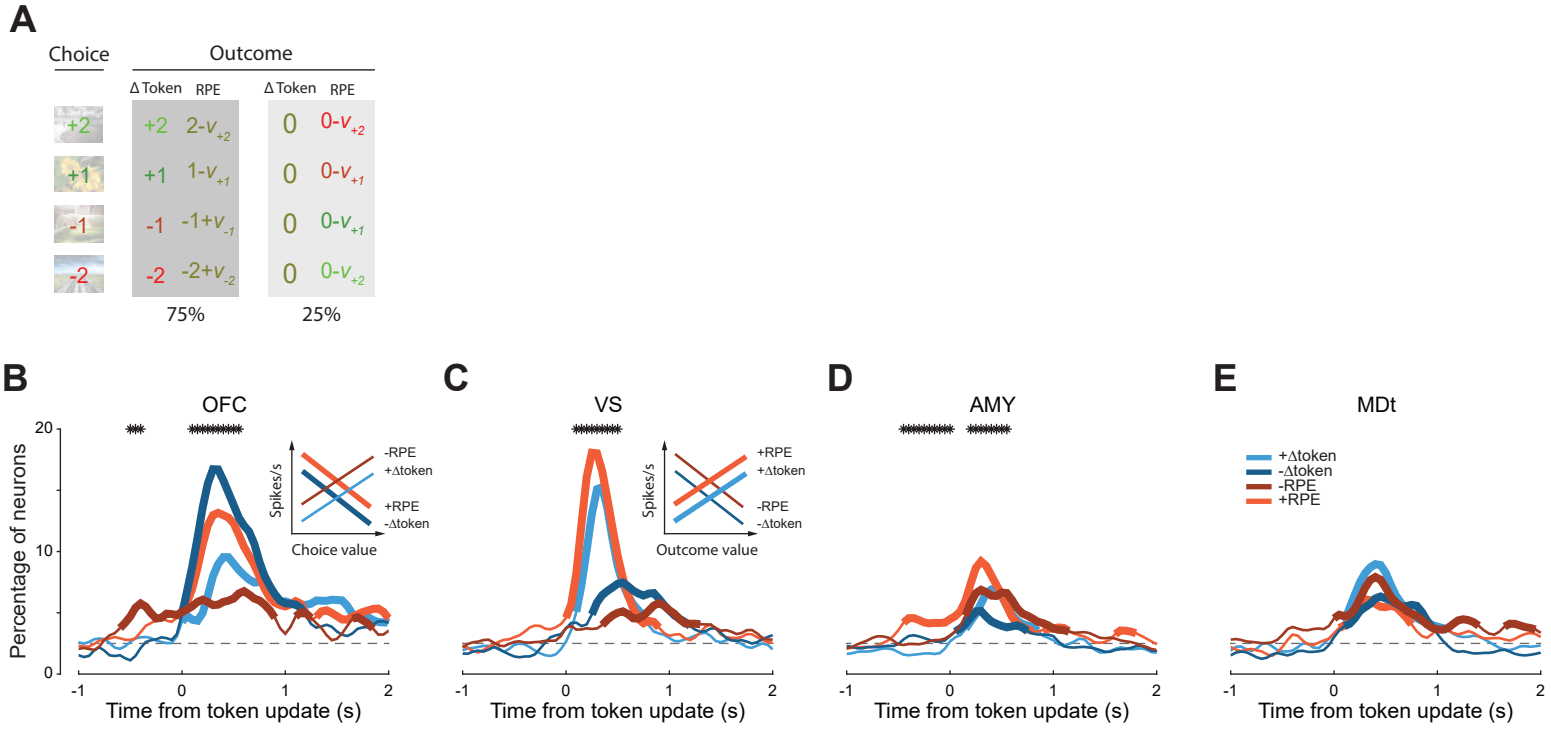
A
Choice
Outcome
Δ Token
RPE
Δ Token
RPE
+2
+2
2-v+2
0
0-v+2
+1
+1
1-v+1
0
0-v+1
-1
-1
-1+v-1
0
0-v+1
-2
-2
-2+v-2
0
0-v+2
75%
25%
B
OFC
Percentage of neurons
Time from token update (s)
Spikes/s
Choice value
-RPE
+Δtoken
+RPE
-Δtoken
C
VS
Time from token update (s)
Spikes/s
Outcome value
+RPE
+Δtoken
-RPE
-Δtoken
D
AMY
Time from token update (s)
E
MDt
Time from token update (s)
+Δtoken
-Δtoken
-RPE
+RPE

Figure 5

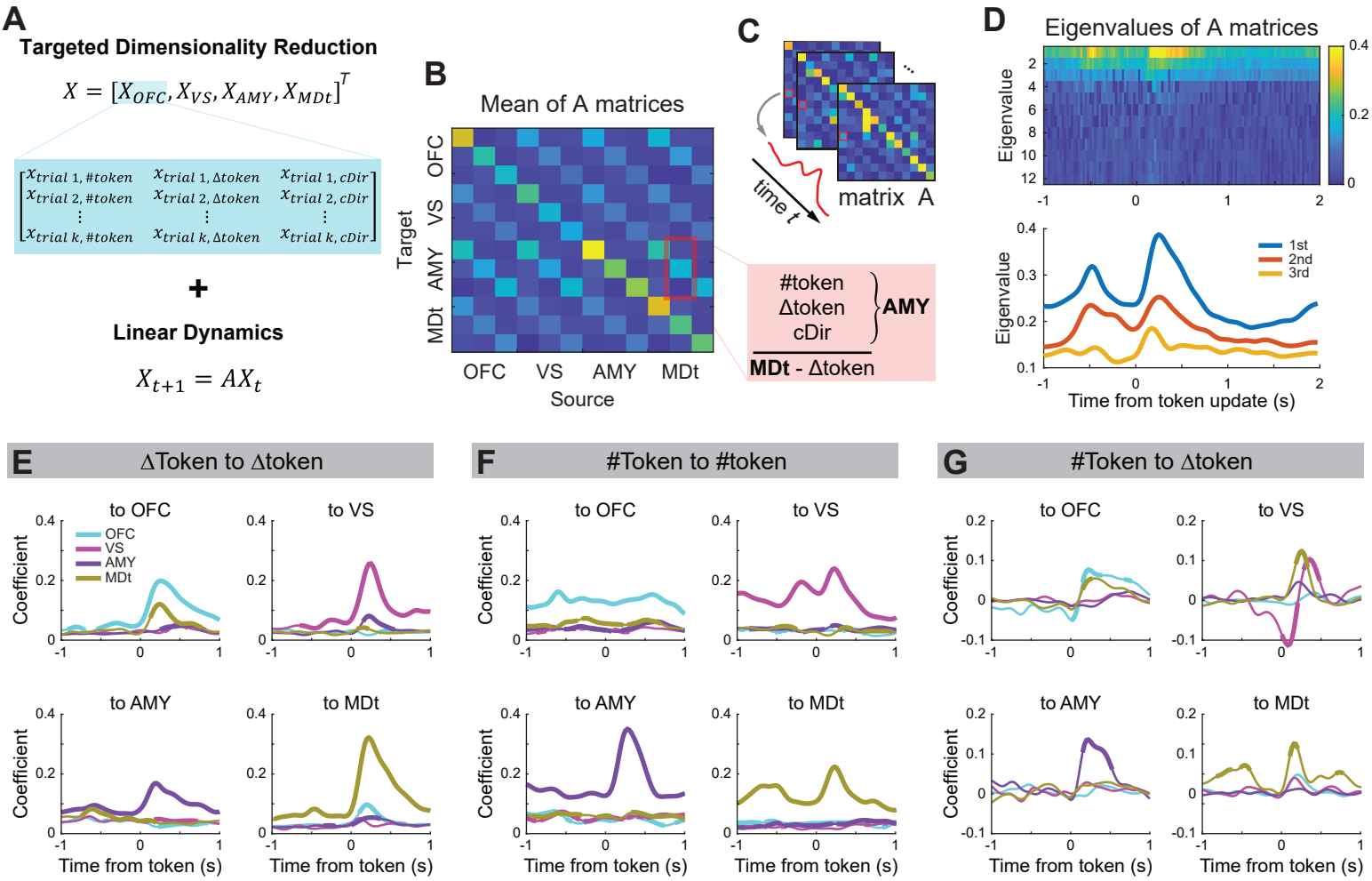
A
Targeted Dimensionality Reduction
$X = [X_{OFC}, X_{VS}, X_{AMY}, X_{MDt}]^T$
$\begin{bmatrix} x_{trial\ 1,\#token} & x_{trial\ 1,\Delta token} & x_{trial\ 1,cDir} \\ x_{trial\ 2,\#token} & x_{trial\ 2,\Delta token} & x_{trial\ 2,cDir} \\ \vdots & \vdots & \vdots \\ x_{trial\ k,\#token} & x_{trial\ k,\Delta token} & x_{trial\ k,cDir} \end{bmatrix}$
+
Linear Dynamics
$X_{t+1} = AX_t$
B
Mean of A matrices
Target
OFC
VS
AMY
MDt
Source
C
time t
matrix A
#token
Δtoken
cDir
AMY
MDt - Δtoken
D
Eigenvalues of A matrices
Eigenvalue
Time from token update (s)
1st
2nd
3rd
E
ΔToken to Δtoken
F
#Token to #token
G
#Token to Δtoken
to OFC
to VS
to AMY
to MDt
Coefficient
Time from token (s)

Figure 6