Feature Specific Prediction Errors and Surprise across Macaque Fronto-Striatal Circuits during Attention and Learning

Mariann Oemisch[1,2], Stephanie Westendorff[1,3], Marzyeh Azimi[1], Seyed Ali Hassani[1,6]; Salva Ardid[4], Paul Tiesinga[5], Thilo Womelsdorf[1,6]

[1]*Department of Biology, Centre for Vision Research, York University, 4700 Keele Street, Toronto, Ontario M6J 1P3, Canada.*

[2]*Department of Neurobiology, Yale University School of Medicine, New Haven, CT 06510.*

[3]*Institute of Neurobiology, University of Tübingen, 72076 Tübingen, Germany.*

[4]*Center for Computational Neuroscience and Neural Technology (CompNet), Department of Mathematics and Statistics, Boston University, Boston, Massachusetts 02215.*

[5]*Donders Institute for Brain, Cognition and Behaviour Radboud University Nijmegen, 6525 EN Nijmegen, Netherlands.*

[6]*Department of Psychology, Vanderbilt University, Nashville, TN 37240*

Abbreviated Title: Feature specific prediction errors across fronto-striatal circuits

Corresponding Authors: Dr. Mariann Oemisch (mariann.oemisch@yale.edu), Dr. Thilo Womelsdorf (thilo.womelsdorf@vanderbilt.edu)

Highlights

• Neural reward prediction errors carry information for updating feature-based attention in all areas of the fronto-striatal network.

• Feature specific neural prediction errors emerge earliest in anterior cingulate cortex and later in lateral prefrontal cortex.

• Ventral striatum neurons encode feature specific surprise strongest for the goal-relevant feature.

• Neurons encoding feature-specific prediction errors contribute to attentional selection after learning.

Summary

Prediction errors signal unexpected outcomes indicating that expectations need to be adjusted. For adjusting expectations efficiently prediction errors need to be associated with the precise features that gave rise to the unexpected outcome. For many visual tasks this credit assignment proceeds in a multidimensional feature space that makes it ambiguous which object defining features are relevant. Here, we report of a potential solution by showing that neurons in all areas of the medial and lateral fronto-striatal networks encode prediction errors that are specific to separate features of attended multidimensional stimuli, with the most ubiquitous prediction error occurring for the reward relevant features. These feature specific prediction error signals (1) are different from a non-specific prediction error signal, (2) arise earliest in the anterior cingulate cortex and later in lateral prefrontal cortex, caudate and ventral striatum, and (3) contribute to feature-based stimulus selection after learning. These findings provide strong evidence for a widely-distributed feature-based eligibility trace that can be used to update synaptic weights for improved feature-based attention.

Introduction

When faced with novel objects we have to learn which features defining these objects are relevant, and which can be safely ignored. To succeed learning which visual features of objects are relevant likely depends on estimating feature relevance and improving this estimate through trial and error learning (Farashahi et al., 2017; Hikosaka et al., 2017; Wilson and Niv, 2011). Computational work shows that this improvement of estimated feature relevance can be achieved by calculating how unexpected an experienced outcome is, and updating its estimate in proportion to this unexpectedness (Leong et al., 2017; Niv et al., 2015). In typical reinforcement learning models, the unexpectedness is calculated as prediction error between predicted and experienced outcome value (Sutton and Barto, 1998; Watkins and Dayan, 1992).

For prediction errors to be useful they need to inform the subject about the specific feature causing the unexpected outcome. In fact, a prominent hypothesis suggests that the degree of

unexpectedness and strength of prediction errors are guiding subject's attention towards those features that gave rise to the unexpected outcome (Gottlieb, 2012; Pearce and Hall, 1980). Biasing attention to those features causing outcomes that have been most unexpected can optimize attentional sampling in the long run to those stimuli with the most reward-predictive features (Daddaoua et al., 2016; Dayan et al., 2000; Ghazizadeh et al., 2016). The mechanisms underlying this attentional optimization through reinforcement learning have been explored in recent studies suggesting that attentional guidance by prediction errors is facilitated when value predictions are already biased towards those feature dimensions that are most likely reward predictive (Hassani et al., 2017; Leong et al., 2017; Niv et al., 2015; Wilson and Niv, 2011). Instead of attending all possible feature dimensions of a stimulus equally, prioritizing those dimensions that most prominently are reward predictive dramatically enhances the learning speed (Farashahi et al., 2017; Kruschke and Hullinger, 2010). This learning speed increase is particularly prominent with stimuli that are defined by multiple dimensions as is typical of real world objects. A strong prediction of these behavioral models is that brain circuits need to combine information about the occurrence of a prediction error with information about the task relevant feature dimension that should be attended (Asaad et al., 2017). However, it is unknown how this combination of prediction error information and feature-based attention is realized in brain circuits.

Here, we set out to identify how this combination of prediction errors and task relevant stimulus features is encoded in medial and lateral fronto-striatal circuits composed of anterior cingulate and ventral striatum as the medial loop and the lateral PFC and caudate nucleus as the lateral loop (Haber and Knutson, 2010). In one scenario of learning which stimulus features are relevant, a general prediction error signal emerges locally within the ventral striatum and is broadcasted to prefrontal cortex where it modifies the activity of feature selective neurons(Asaad et al., 2017). Updated prefrontal cortex neurons might then exert an improved top-down signal over sensory cortices for attention and choices in subsequent trials (e.g. (Fusi et al., 2007; Seo et al., 2012). This view is supported by a ubiquity of mostly human fMRI findings that single out the striatum as core region to encode prediction errors (Chase et al., 2015; Glimcher, 2011), and the lateral prefrontal cortex to encode a feature-based top-down signal (Leong et al., 2017; Serences, 2008) together with prediction error information (Asaad et al., 2017). In contrast to such a view emphasizing functional localization, neurons encoding prediction errors in multiple areas might carry already feature information. Such feature specific prediction errors could serve as feature

specific eligibility trace orchestrated across the recurrent fronto-striatal loops. Such a distributed, feature-specific eligibility trace is predicted by spiking network models that learn task relevant features by using attentional feedback signals to label synapses among those neurons that also contributed to the feature specific reward prediction itself (Roelfsema and van Ooyen, 2005; Rombouts et al., 2015).

Here, we tested these scenarios by recording from four brain areas of the medial and lateral fronto-striatal loops in macaque monkeys performing a feature-based value learning task. Task performance was well fit by an attention-weighting reinforcement learning model that estimated trial-by-trial prediction errors. We found that substantial proportions of neurons across all brain areas encoded prediction errors for task relevant features. The strongest encoding was evident for the task and reward relevant feature, which represents a versatile eligibility trace that can guide synaptic learning across the whole fronto-striatal loop. We found that this feature specific eligibility trace emerged after an initial unspecific prediction error signal and was associated with stronger attentional selection signals in subsequent trials, thus potentially contributing to improved learning and visual selection.

Results

Behavior

Monkeys performed a reversal learning task which presented in each trial two peripheral stimuli with different colors and motion directions (**Figure 1A**). Over sequences of 30 or more correct trials one of two colors was associated with reward outcomes (juice drops), while no other feature (left vs. right stimulus location, or up vs. downward motion direction, or the alternative color) was linked to reward (**Figure 1B**). In order to obtain reward the animals had to wait for a Go-signal (dimming of the stimuli) and make a saccade in the motion direction of the grating stimulus whose color matched the reward associated color. This task required (1) feature-based attentional selection of one over another stimulus based on a reward associated color, and (2) to use the motion direction of the attended stimulus to program a saccadic response. Above chance performance on this task required learning that a correct (rewarded) or incorrect (nonrewarded) outcome was caused by the attended color of the stimulus rather than by its motion direction or its location on the screen. Both monkeys learned this feature specific credit assignment and adjusted their feature-

based attention bias to the reward associated color after uncued reversals (**Figure 1C**). As estimated with an ideal observer statistic (Balcarras et al., 2016; Smith et al., 2004), monkey H / K successfully learned an average of 83 / 91% of blocks, whereby learning occurred on average within 17.5 / 16.5 number of trials following the reversal. Monkey H / K performed an average of 9.7±0.3 / 8.9±0.3 reversal blocks per recording session with an average block length of 46±0.7 (*monkey H,* median = 37) and 43±0.8 (*monkey K*, median = 36) trials.

Neuronal Encoding of Outcome and Feature-Specific Reward Prediction Errors

During reversal learning performance, we recorded 1960 units in two monkeys with 690 units in the ACC (*monkey H*: 405, *monkey K*: 285), 524 units in PFC (*monkey H*: 316, *monkey K*: 208), 449 units in caudate nucleus (CD; *monkey H*: 234, *monkey K*: 215) and 297 units in ventral striatum (VS; *monkey H*: 163, *monkey K*: 134) (**Figure 1E**). 71% of neurons in *monkey H* and 78% of neurons in *monkey K* met the criteria for analysis (*see* Methods). Among these neurons 38% encoded the outcome (rewarded versus unrewarded), ranging between 27-53% across brain areas (**Suppl. Figure S1A, S3A, B**). To discern trial-by-trial encoding of prediction errors we fitted an attention weighting reinforcement learning model to the choice data of the monkeys (**Figure 1D**, *see* Methods) (Hassani et al., 2017; Leong et al., 2017; Wilson and Niv, 2011). We then correlated the model derived reward prediction errors (RPEs) with the neural firing rates during the 0-1.5 sec. reward outcome epoch. For monkeys H/K the firing rates correlated significantly with positive RPE (following correct trials) in 21/22% of neurons, with the negative RPE (following incorrect trials) in 14/10% of neurons, and for 24/24% of neurons with the unsigned RPE that indexes surprise (e.g. (Hayden et al., 2011)). Firing correlations with predictions errors were evident in each of the recorded brain areas in both monkeys (**Suppl. Figure S1B**).

We next hypothesized that in order to use prediction error information to adjust feature-based attention, neurons might encode RPEs not equally for all features of the stimulus, but selectively for the task relevant features. We found support for this suggestion in multiple example neurons with such feature selective encoding of RPE (**Figure 2**). For instance, the VS neuron in **Figure 2i** scales its firing rate with surprise (unsigned RPE) when color 1 was selected for the choice (top), but not when color 2 was selected (middle). The ACC neuron in **Figure 2iv** scales its firing rate with the negative RPE (greater firing with more negative RPE) when the selected

stimulus was located on the left (top), but not when it was located on the right (middle). And the PFC neuron in **Figure 2v** scales its firing rate with the positive RPE, when the stimulus selected in the preceding choice was located on the right (top), but not when it was located on the left (middle). Overall, we found that 53.1% of neurons (*monkey H*: 52.7%, *monkey K*: 53.6%) across the fronto-striatal areas tested here encoded feature-specific positive, negative and unsigned surprise signals (**Suppl. Figures S6, S7**).

Feature-specific RPE Signals Emerge Later than Non-specific RPEs and Earliest in ACC

Feature specific RPE signals might arise from neurons that initially encode the occurrence of a non-specific prediction error by combining RPE with feature information over time. This suggestion predicts a slower time course of more specific information about the source of the error (Schultz, 2016). We tested this possibility by determining for each neuron the time window in which it significantly encoded a feature-specific RPE, non-specific RPE, or outcome per se for four consecutive time bins (≥0.1 sec.), and comparing their average time-courses (*see* Methods). We found across neurons that feature specific RPE encoding emerged significantly slower than non-specific RPE encoding as indexed by a shallower slope of the temporal cumulation of the proportion of significantly RPE encoding neurons (Kolmogorov-Smirnoff test, Bonferroni-Holm multiple comparison corrected: $p_{feat\text{-}non} < .001$) (**Figure 3A, B**). Prediction errors in general emerged significantly later than the encoding of the rewarded/nonrewarded outcome (Kolmogorov-Smirnoff test, Bonferroni-Holm multiple comparison corrected: $p_{feat\text{-}non} < .001$; $p_{feat\text{-}out} < .001$; $p_{non\text{-}out} < .001$). In addition, feature specific RPE encoding continued to increase and remained at a higher plateau level for a longer duration than nonspecific RPE signals (**Figure 3A**). Across units, 25% of outcome encoding occurred at 268ms, while 25% of non-specific RPE encoding occurred 283ms after feedback onset, and 25% of feature-specific RPE encoding occurred after 355ms (Randomization statistic: $p_{feat\text{-}non} < .001$; $p_{feat\text{-}out} < .001$; $p_{non\text{-}out} = .008$; **Figure 3B**).

We next ask when feature specific RPE encoding emerges in each of the four brain areas. Using the same latency measures as above, we found that the rise of neurons with significant feature specific RPE differed significantly between all areas, except for ACC and CD which did not differ (Kolmogorov-Smirnoff test, Bonferroni-Holm multiple comparison corrected: $p_{ACC\text{-}PFC} < .001$; $p_{ACC\text{-}CD} = .128$; $p_{ACC\text{-}VS} < .001$; $p_{PFC\text{-}CD} < .001$; $p_{PFC\text{-}VS} = .006$; $p_{CD\text{-}VS} < .001$) (**Figure 4A,**

B). Feature-specific RPE signals emerged earliest in the ACC (310ms) and CD (330ms), followed by PFC (385ms), followed by VS (428ms) (Randomization statistic: $p_{ACC\text{-}PFC} < .001$; $p_{ACC\text{-}CD} = .136$; $p_{ACC\text{-}VS} < .001$; $p_{PFC\text{-}CD} < .001$; $p_{PFC\text{-}VS} = .018$; $p_{CD\text{-}VS} < .001$; **Figure 4B** bottom).

Feature-tuning of Reward Prediction Errors

To guide reversal learning towards the goal-relevant color feature, subgroups of neurons should preferentially encode the prediction error for the reward-relevant color dimension. Such color specific error signals are likely candidates to update the task set representation following a reversal. In contrast to color, motion and location were choice-relevant stimulus dimensions that are important for completing the current trial selection, but do not facilitate reversal learning of the new color-reward rule. Consistent with this rationale, we found negative and positive RPEs were encoded more often for the reward-relevant color dimension, than for the location or motion dimension (one-sided bootstrap CI: $p \leq 0.05$; **Figure 5A and D,** respectively). When split by areas, we found that neurons with color-specific negative RPEs were more prevalent than location- or motion-specific negative RPEs in ACC, VS, and PFC (one-sided bootstrap CI: $p \leq 0.05$, **Figure 5B**). We used an index to quantify the relative proportion of color-selective RPE neuron compared to location- and motion- selective RPEs $[(RPE_{col} - RPE_{loc+motion}/2) / (RPE_{col} + RPE_{loc} + RPE_{motion})]$ with >0 values indicating a preference to encode RPEs for the goal relevant color feature. This color tuning index showed that for negative RPEs ACC, VS, and PFC showed stronger color tuned RPEs than CD (two-sided bootstrap CI: $p \leq 0.05$; **Figure 5C**, see Methods). Similar to negative RPEs, positive RPEs were more often color-specific than location- or motion-specific in ACC and VS (one-sided bootstrap CI: $p \leq 0.05$, **Figure 5E**, left column). In addition, neurons in CD also selectively encoded feature-specific positive RPEs in the color dimension, while neurons in PFC were not selective (**Figure 5E**, right column). Color tuning indices did not differ substantially between areas (ACC: $I_{col} = 0.10$, VS: $I_{col} = 0.14$, PFC: $I_{col} = 0.05$, CD: $I_{col} = 0.123$; two-sided bootstrap CI: $p > 0.05$; **Figure 5F**). The average correlation strengths between RPE and firing rate across time for those neurons that encoded a color-specific negative or positive RPE were comparable across areas (**Suppl. Figure S2**).

In contrast to positive and negative RPEs, unsigned surprise signals were across areas similarly prevalent for the color, location and motion dimensions (one-sided bootstrap CI: $p > 0.05$; **Figure 6A**). Split by areas, only the ventral striatum encoded surprise signals stronger for

color than motion and location (one-sided bootstrap CI: $p \leq 0.05$, **Figure 6B** bottom left). This finding was confirmed by a significantly higher color tuning index for VS ($I_{col} = 0.103$) than for ACC ($I_{col} = -0.024$), PFC ($I_{col} = -0.036$), and CD ($I_{col} = -0.01$) (**Figure 6C**). The average correlation strength of color-specific unsigned RPE units was similar across areas (**Suppl. Figure S2**).

Feature-Specific RPEs are Largely Segregated from Feature-Specific Outcome Signals

Feature specific correlations of firing rates with the RPE signals might be evident in populations of neurons that show already feature specific firing for the outcome itself irrespective of prediction error, or they could occur in a segregated neuronal population. To answer this question, we first calculated the prevalence of feature information in the outcome period. We found that 20/16/13% of neurons encoded color/motion/location specific information in the outcome epoch at the first or second order (see **Methods** and **Suppl. Fig. S3**). However, only 35/28/26/% of these neurons also correlated their firing with color/motion/location specific RPEs, suggesting that a substantial population of feature-specific RPE encoding neurons cannot be explained by error independent feature preferences (**Suppl. Fig. S4**).

Cell-type Specificity of RPE Encoding Neurons

To understand the mechanisms underlying feature specific prediction errors it will be important to identify the functional cell types encoding them. In our recordings, we used the action potential waveforms to distinguish two functional cell types in the cortical brain areas (putative pyramidal cells and putative interneurons), and two cell types in the striatum (putative medium spiny neurons and putative interneurons) using methods established before (Ardid et al., 2015; Berke, 2008; Lansink et al., 2010) (see **Methods and Fig 7**). In the cortical areas ACC and PFC, we found that narrow spiking neurons (putative interneurons) were more likely to encode feature-specific RPE signals (ratio narrow/broad = 0.65) than we would expect based on the population distribution of narrow to broad spiking units we recorded (ratio narrow/broad in population = 0.40) (Chi-square test, $p \leq .05$), and that this was not the case for neurons encoding non-specific RPE signals (ratio narrow/broad = 0.53; Chi-square test, $p > .05$; **Fig 7C-E)**. The same effect was visible as a statistical trend for the striatal areas, caudate and ventral striatum, whereby feature-specific RPEs were more frequently encoded by narrow spiking neurons (putative interneurons, (Berke et al.,

2004; Kawaguchi, 1993; Lansink et al., 2010) than suggested based on the population distribution (Chi-square test feature-specific RPEs: p = .08) (**Fig 7H-J).**

Neuronal Signaling of Feature-Specific Prediction Errors can affect Stimulus Selection

What are the functional consequences of feature specific prediction errors? At the behavioral level, prediction errors indicate the need to adjust attention in subsequent trials. At the neuronal level, this adjustment for future attention might correspond to a shift of firing from the outcome time epoch early during learning to the firing at the time epoch of stimulus selection after learning. This temporal transfer of firing is the classical signature of reward prediction error encoding by ventral tegmental dopaminergic neurons (Fiorillo et al., 2003; Schultz, 1998; Schultz et al., 1993). To test whether such a transfer takes place for feature specific prediction errors, we determined whether the magnitude of the prediction error in the current trial was related to firing rate changes during stimulus selection in the following trial. We hypothesized that during learning periods when prediction errors are large, neurons would not yet contribute to the visual selection of the stimulus, but after learning when prediction errors are low, the same neurons would show an enhanced stimulus onset response indicating that they contribute to the attentional selection of the stimulus. For each color-specific RPE encoding neuron, we found the 25% of trials with the largest RPE and the 25% of trials with the smallest RPE, and compared the change in firing rate from pre- to post-color onset in those trials. On average, across these color-specific RPE encoding neurons we found an increased firing from pre- to post-color onset (t-test, $p < .0001$ for each RPE type). This increased color onset response was on average (1) stronger following trials with low RPE than high RPE and (**Figure 8A, B** lower versus upper histograms, and **Suppl. Fig S5**), and (2) stronger when the next trials choice was for the preferred color of the neurons (**Figure 8A, B** cyan versus grey histograms) The difference in firing rate change for low versus high RPE trials was statistically significant for neurons encoding positive RPE (paired t-test, $p < .001$; $p_{nonpref} = .185$) (**Figure 8A**), and for neurons encoding surprise (paired t-test, $p < .001$, $p_{nonpref} = .065$) (**Figure 8B**), but not for neurons encoding negative RPE ($p = .089$, $p_{nonpref} = .291$; data not shown).

The selectively increased firing to color onset after low RPE trials was most prominent and statistically significant for ACC neurons encoding color specific positive RPEs (**Fig 8C, Suppl. Fig S5**), and for CD neurons encoding color specific surprise (**Figure 8D, Suppl. Fig S5**). These finding provide strong evidence that feature specific prediction error signals during learning

translate into color cue firing rate increases for these same neurons after learning has taken place, reminiscent of the temporal transfer of classical dopaminergic prediction error signals.

Discussion

We found that about half of the neuronal populations in anterior cingulate cortex, lateral prefrontal cortex, ventral striatum and caudate nucleus encoded reward prediction errors specifically for one of three features of an attended visual stimulus during value learning. This feature specific prediction error was significantly stronger for the task relevant feature that also predicted reward over trials in a block (color), indicating that feature specific prediction error signals carry goal-relevant information that can bias improved feature-based attention in future trials. Feature specific prediction error encoding emerged in time after nonspecific prediction error encoding, indicating that it is based on a partly independent process that combines prediction error information with feature information over time.

Prediction errors carry goal-relevant information to adjust feature-based attention

Among all recorded brain areas, the anterior cingulate cortex stood out by containing most neurons with early feature specific prediction error information, with a slower rise of feature information in the caudate head, lPFC, and ventral striatum (**Figure 4B**). This finding underlines the importance of the anterior cingulate cortex to indicate the specific information needed to adjust behavior in future trials (Shen et al., 2015; Shenhav et al., 2013; Shenhav et al., 2016). In our task, the goal-information was the specific color of the attended stimulus. This finding complements previous reports of anterior cingulate cortex neurons conveying prediction error related activity for specific actions (Matsumoto et al., 2007; Quilodran et al., 2008), unique objects (Kennerley et al., 2009), stimulus-response mapping rules (Johnston et al., 2007; Womelsdorf et al., 2010), attentional and motivational origin of errors (Shen et al., 2015), and more abstract combinations of stimulus and reward information (Kennerley et al., 2011). Our ACC finding uniquely adds to this literature by showing firstly, that ACC prediction error activity is combined with the attended color feature in an attention task that always presents all possible features on the screen, which induces perceptual ambiguity and uncertainty about stimulus features (**Figure 1**). Secondly, that

this feature-specific prediction error activity potentially results in greater feature-based attention activity selectively for the preferred color with learning (**Figure 8**). These findings complement reports of attention specific activity in the ACC (Kaping et al., 2011; Oemisch et al., 2015; Voloh et al., 2015) supporting the view that ACC plays a major role for controlling to which of the available options (covert) attention shifts (Mesulam, 1981; Westendorff et al., 2016; Womelsdorf and Everling, 2015).

Thirdly, our findings clarify that specific information about the origin of prediction errors is not localized to the ACC, but widely distributes to all areas we recorded from and which are known to be anatomically mono-synaptically connected (Barbas and Pandya, 1989; Haber and Knutson, 2010; Hikosaka et al., 2017; Medalla and Barbas, 2009; Morecraft et al., 1993; Morecraft et al., 2012), and functionally synchronized in different task contexts (Antzoulatos and Miller, 2014; Oemisch et al., 2015; Voloh et al., 2015; Womelsdorf et al., 2014). The distributed nature of prediction error signaling supports recent summaries of the recurrent nature of fronto-striatal processes underlying reward-based choices (Hunt and Hayden, 2017), but also illustrates that latency analysis is able to identify a single brain area such as the ACC to have a particularly early functional contribution to value based attention and learning in a demanding reversal learning task used here.

The preponderance of narrow spiking neurons (in cortex) and as a trend of narrow spiking neurons in striatum to carry feature-specific error information provided an unexpected, data driven finding that support suggestions of a particular role of inhibitory neurons to process learning related information and/or to induce plasticity in cortical and striatal networks (Berke, 2011; Hennequin et al., 2017; Vogels et al., 2011). Previous studies suggest that the action potential waveform corresponds to inhibitory neurons in cortex and striatum (Kawaguchi, 1993; Plenz and Kitai, 1998; Wilson et al., 1994). Assuming that our finding indicate that putative interneurons are particularly informative about the error term is consistent with their involvement to regulate network level plasticity changes (Hennequin et al., 2017). Previous work has shown for instance that spike timing dependent plasticity in rodent corticostriatal excitatory synapses is crucially dependent on GABAergic signaling of inhibitory circuits (Paille et al., 2013), and that in balanced networks changes in inhibitory synaptic strength are accompanying changes in excitatory synaptic changes (Vogels et al., 2011).

A role of surprise signals in the ventral striatum to guide 'attention for learning'

For the ACC, PFC, CD and VS, prediction errors correlated with the firing of neurons after correct trials giving rise to positive RPEs, after incorrect trials giving rise to negative RPEs, and independent of the actual trial outcomes giving rise to unsigned unexpectedness, or surprise. Large surprise signals (to rare, high rewards) in the ACC have previously been shown to predict adjustment of behavioral strategies (Hayden et al., 2011), but it has been questioned whether any surprise related activity exists that relates to changes in attention (Le Pelley et al., 2016). Here, we found widely distributed and prevalent, neuronal surprise signals carrying significant information about all features of an attended stimulus to a similar degree in ACC, PFC, and CD. A notable exception was the ventral striatum which contained proportionally stronger neuronal surprise signals for the goal relevant color feature as opposed to the task relevant, but reward irrelevant, location and motion feature (**Figure 6C**). This preponderance of goal-relevant feature information in the ventral striatum is particularly noteworthy in light of a long-standing psychological theory of attention suggesting that attention during learning is driven by unexpected events (including outcome events) (Gottlieb, 2012, 2017; Pearce and Hall, 1980). According to this attention model, unexpected outcomes should give rise to stronger visual selection of the stimulus feature that gave rise to the violated expectation (Gottlieb et al., 2014). Our study directly tested this hypothesis and confirmed that the same neural population that encodes the feature specific surprise also showed stronger firing rate increases after the feature onset in subsequent trials (**Figure 8**). This increased feature selection signal (1) was stronger for the color that was preferred versus non-preferred by the neuron, and (2) it was stronger when subjects had learned the relevant feature, i.e. when prediction errors were comparably low in previous trials. These results provide strong evidence for a role of the ventral striatum to directly contribute to the learning of and attentional biasing towards goal relevant features. This conclusion adds an important functional role to prediction error signaling which is - across species - ubiquitously reported to be particularly strong in the ventral striatum (Chase et al., 2015; Cockburn et al., 2014; Costa et al., 2016; Diederen et al., 2016; Leong et al., 2017; O'Doherty et al., 2017; Schultz, 2017; Schultz et al., 2003; Takahashi et al., 2016; Watabe-Uchida et al., 2017).

Additionally, the finding of feature encoding in the ventral striatum and the caudate during learning of feature-based attentional allocation supports recent re-conceptualization of attentional biases as reflecting the internal striatal activity state that resolved competing value predictions and

beliefs about possible relevant stimuli (Krauzlis et al., 2014; Womelsdorf and Everling, 2015). In these hypotheses, attention is not considered to reflect a unitary top-down signal that is obscure and localized to the prefrontal cortex as in many classical models, but rather attention emerges from (is the effect of) the current state of basal ganglia circuits that continuously integrate multiple information streams and resolves competition among these inputs (Krauzlis et al., 2014). A core insight from this hypothesis is that the striatum has direct access to the spatial maps in the superior colliculus via disinhibitory circuits in the substantia nigra (e.g. (Hikosaka et al., 2017)). With this direct access to the superior colliculus, activity in ventral striatum and caudate nucleus exert a direct bias for overt fixational sampling and covert attentional selection of visual information (Ignashchenkova et al., 2004; Lovejoy and Krauzlis, 2010; Zenon and Krauzlis, 2012). The results of our study support these hypotheses by revealing widespread feature specific prediction errors (**Figures 5 and 6**) and feature specific selection effects (**Figures 8** and **Supplementary Figure S5**) across the medial fronto-striatal loop (ACC and ventral striatum) and the lateral fronto-striatal loops (rostralateral PFC and dorsal caudate head).

Neuronal credit assignment for relevant features with multidimensional stimuli

We report of neuronal populations encoding the 'goal-specific' reward prediction error for one color, but not for another color. These neuronal groups encode how unexpected it was that the attended color led to a reward, or to omission for reward. This information is precisely what is needed to enhance those synaptic connections between neurons that encode the specific color that is more relevant than expected, and to reduce the synaptic connection weights among neurons encoding the color that was less rewarded than expected. These types of synaptic weight changes following the strength of prediction errors have been successfully modelled in several spiking network models implementing reinforcement learning using different synaptic learning rules (Fremaux et al., 2013; Friedrich et al., 2011; Fusi et al., 2007; Potjans et al., 2011; Rasmussen et al., 2017; Rombouts et al., 2015; Santoro et al., 2016; Seung, 2003; Soltani and Wang, 2010; Suri and Schultz, 1998; Urbanczik and Senn, 2009). These models illustrate, for example, that simpler stimulus-response reversal learning performance in monkeys can be realized by spike timing dependent plasticity changes (Fusi et al., 2007). However, it has remained unclear how to implement more complex credit assignment in a higher dimensional feature space where multiple features could be credited for an outcome, even though only one feature is actually relevant (Niv

et al., 2015; Wilson and Niv, 2011). For this situation, a recent spiking model suggested a 4-factor learning rule that is dependent on attention to a specific stimulus feature or action prior to registering a reward/no-reward outcome (Rombouts et al., 2015). In this model, neurons activated by an outcome receive a synaptic tag, which is specific to the attended feature, from feedback connections originating from output neurons similar to striatal output neurons. This attentional feedback induced synaptic tag acts like an attention specific eligibility trace that can be combined with dopamine dependent (feature unspecific) prediction error information when a (rewarding or non-rewarding) outcome is received. Learning is achieved when these two factors (attentional feedback and neuromodulatory prediction error information) meet at the synapses between pairs of neurons that showed near coincident pre- and postsynaptic activity during the outcome processing (Rombouts et al., 2015). The models make multiple predictions that are supported by our data. Firstly, synaptic updating is taking place in an associative network layer that resembles the fronto-striatal network of value learning as opposed to sensory or motor related network layers. Secondly, feature specific prediction errors are predicted to emerge as local neuronal signals across the entire associative network based on neuron-specific synaptic tags, closely corresponding to the distributed RPEs we observed. Finally, the model predicts that a learning of task relevant features depends on attention towards those stimulus features that are most consistently reward associated. This attentional hypothesis of reinforcement learning was directly tested in our experiment, providing evidence that the most ubiquitously encoded prediction error signals occur for the attended, goal-relevant color feature.

Taking together, we believe that our findings provide direct support of the concept of attention weighted reinforcement learning as a generic framework to understand learning and attention in environments with multidimensional stimuli, as is typical for real life learning of object relevance (Hassani et al., 2017; Leong et al., 2017; Niv et al., 2015). The existence of network-wide available information about the degree to which individual features of visual stimuli led to unexpected outcomes will further constrain the modeling of learning rules that efficiently solve the credit assignment problem (Farashahi et al., 2017; Hennequin et al., 2017; Watabe-Uchida et al., 2017). Our study may provide a starting point documenting that network wide credit assignment processes are directly related to improved biases of feature-based visual attention. It will be an important question for future research to specify how feature specific eligibility traces are used to determine the value of objects during fast and slow learning.

Acknowledgement: This work was supported by grant MOP 102482 from the Canadian Institutes of Health Research (TW) and by the Natural Sciences and Engineering Research Council of Canada (TW), and the Brain in Action CREATE-IRTG program (MO, TW). The funders had no role in study design, data collection and analysis, the decision to publish, or the preparation of this manuscript. Authors would like to thank Hongying Wang, Samira Azimi and Ali Hassani for technical support.

Author contributions: Conceptualization, T.W.; Methodology, T.W., S.W. and M.O.; Formal Analysis, M.O., T.W., S.A. and P.T; Investigation, M.O., M.A., S.A.H.; Writing – Original draft, M.O. and T.W.; Writing – Review & Editing, all authors; Supervision, T.W.; Funding acquisition, T.W.

Declaration of interests: The authors declare no competing interests.

References

Antzoulatos, E.G., and Miller, E.K. (2014). Increases in Functional Connectivity between Prefrontal Cortex and Striatum during Category Learning. Neuron 83, 216-225.
Ardid, S., Vinck, M., Kaping, D., Marquez, S., Everling, S., and Womelsdorf, T. (2015). Mapping of functionally characterized cell classes onto canonical circuit operations in primate prefrontal cortex. Journal of Neuroscience 35 2975–2991.
Asaad, W.F., Lauro, P.M., Perge, J.A., and Eskandar, E.N. (2017). Prefrontal Neurons Encode a Solution to the Credit-Assignment Problem. J Neurosci 37, 6995-7007.
Balcarras, M., Ardid, S., Kaping, D., Everling, S., and Womelsdorf, T. (2016). Attentional Selection Can Be Predicted by Reinforcement Learning of Task-relevant Stimulus Features Weighted by Value-independent Stickiness. J Cogn Neurosci 28, 333-349.
Barbas, H., and Pandya, D.N. (1989). Architecture and intrinsic connections of the prefrontal cortex in the rhesus monkey. The Journal of comparative neurology 286, 353-375.
Berke, J.D. (2008). Uncoordinated firing rate changes of striatal fast-spiking interneurons during behavioral task performance. The Journal of neuroscience : the official journal of the Society for Neuroscience 28, 10075-10080.
Berke, J.D. (2011). Functional properties of striatal fast-spiking interneurons. Front Syst Neurosci 5, 45.
Berke, J.D., Okatan, M., Skurski, J., and Eichenbaum, H.B. (2004). Oscillatory entrainment of striatal neurons in freely moving rats. Neuron 43, 883-896.
Cai, X., and Padoa-Schioppa, C. (2014). Contributions of Orbitofrontal and Lateral Prefrontal Cortices to Economic Choice and the Good-to-Action Transformation. Neuron 81, 1140-1151.
Calabrese, E., Badea, A., Coe, C.L., Lubach, G.R., Shi, Y., Styner, M.A., and Johnson, G.A. (2015). A diffusion tensor MRI atlas of the postmortem rhesus macaque brain. Neuroimage 117, 408-416.
Chase, H.W., Kumar, P., Eickhoff, S.B., and Dombrovski, A.Y. (2015). Reinforcement learning models and their neural correlates: An activation likelihood estimation meta-analysis. Cognitive, affective & behavioral neuroscience 15, 435-459.
Cockburn, J., Collins, A.G., and Frank, M.J. (2014). A reinforcement learning mechanism responsible for the valuation of free choice. Neuron 83, 551-557.
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.
Daddaoua, N., Lopes, M., and Gottlieb, J. (2016). Intrinsically motivated oculomotor exploration guided by uncertainty reduction and conditioned reinforcement in non-human primates. Scientific reports 6, 20202.
Dayan, P., Kakade, S., and Montague, P.R. (2000). Learning and selective attention. Nature neuroscience 3 Suppl, 1218-1223.
Diederen, K.M., Spencer, T., Vestergaard, M.D., Fletcher, P.C., and Schultz, W. (2016). Adaptive Prediction Error Coding in the Human Midbrain and Striatum Facilitates Behavioral Adaptation and Learning Efficiency. Neuron 90, 1127-1138.
Donahue, C.H., and Lee, D. (2015). Dynamic routing of task-relevant signals for decision making in dorsolateral prefrontal cortex. Nat Neurosci.
Dunn, O.J., and Clark, V.A. (1987). Applied Statistics: Analysi of Variance and Regression.
Farashahi, S., Rowe, K., Aslami, Z., Lee, D., and Soltani, A. (2017). Feature-based learning improves adaptability without compromising precision. Nat Commun 8, 1768.
Fiorillo, C.D., Tobler, P.N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. Science 299, 1898-1902.
Fremaux, N., Sprekeler, H., and Gerstner, W. (2013). Reinforcement learning using a continuous time actor-critic framework with spiking neurons. PLoS computational biology 9, e1003024.
Friedrich, J., Urbanczik, R., and Senn, W. (2011). Spatio-temporal credit assignment in neuronal population learning. PLoS computational biology 7, e1002092.
Fusi, S., Asaad, W.F., Miller, E.K., and Wang, X.J. (2007). A neural circuit model of flexible sensorimotor mapping: learning and forgetting on multiple timescales. Neuron 54, 319-333.

Genovesio, A., Tsujimoto, S., Navarra, G., Falcone, R., and Wise, S.P. (2014). Autonomous encoding of irrelevant goals and outcomes by prefrontal cortex neurons. J Neurosci 34, 1970-1978.
Ghazizadeh, A., Griggs, W., and Hikosaka, O. (2016). Ecological Origins of Object Salience: Reward, Uncertainty, Aversiveness, and Novelty. Frontiers in Neuroscience 10.
Glantz, S., and Slinker, B. (2001). Primer of Applied Regression and Analysis of Variance.
Glimcher, P.W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. P Natl Acad Sci USA 108 Suppl 3, 15647-15654.
Gottlieb, J. (2012). Attention, learning, and the value of information. Neuron 76, 281-295.
Gottlieb, J. (2017). Understanding active sampling strategies: Empirical approaches and implications for attention and decision research. Cortex; a journal devoted to the study of the nervous system and behavior.
Gottlieb, J., Hayhoe, M., Hikosaka, O., and Rangel, A. (2014). Attention, reward, and information seeking. J Neurosci 34, 15497-15504.
Haber, S.N., and Knutson, B. (2010). The reward circuit: linking primate anatomy and human imaging. Neuropsychopharmacology 35, 4-26.
Hassani, S.A., Oemisch, M., Balcarras, M., Westendorff, S., Ardid, S., van der Meer, M.A., Tiesinga, P., and Womelsdorf, T. (2017). A computational psychiatry approach identifies how alpha-2A noradrenergic agonist Guanfacine affects feature-based reinforcement learning in the macaque. Scientific reports 7.
Hayden, B.Y., Heilbronner, S.R., Pearson, J.M., and Platt, M.L. (2011). Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. J Neurosci 31, 4178-4187.
Hennequin, G., Agnes, E.J., and Vogels, T.P. (2017). Inhibitory Plasticity: Balance, Control, and Codependence. Annu Rev Neurosci 40, 557-579.
Hikosaka, O., Ghazizadeh, A., Griggs, W., and Amita, H. (2017). Parallel basal ganglia circuits for decision making. J Neural Transm (Vienna).
Hunt, L.T., and Hayden, B.Y. (2017). A distributed, hierarchical and recurrent framework for reward-based choice. Nature Reviews Neuroscience 18, 172-182.
Ignashchenkova, A., Dicke, P.W., Haarmeier, T., and Thier, P. (2004). Neuron-specific contribution of the superior colliculus to overt and covert shifts of attention. Nat Neurosci 7, 56-64.
Johnston, K., Levin, H.M., Koval, M.J., and Everling, S. (2007). Top-down control-signal dynamics in anterior cingulate and prefrontal cortex neurons following task switching. Neuron 53, 453-462.
Kaping, D., Vinck, M., Hutchison, R.M., Everling, S., and Womelsdorf, T. (2011). Specific contributions of ventromedial, anterior cingulate, and lateral prefrontal cortex for attentional selection and stimulus valuation. PLoS Biology 9, e1001224.
Kawaguchi, Y. (1993). Physiological, morphological, and histochemical characterization of three classes of interneurons in rat neostriatum. J Neurosci 13, 4908-4923.
Kennerley, S.W., Behrens, T.E., and Wallis, J.D. (2011). Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. Nature neuroscience 14, 1581-1589.
Kennerley, S.W., Dahmubed, A.F., Lara, A.H., and Wallis, J.D. (2009). Neurons in the frontal lobe encode the value of multiple decision variables. J Cogn Neurosci 21, 1162-1178.
Krauzlis, R.J., Bollimunta, A., Arcizet, F., and Wang, L. (2014). Attention as an effect not a cause. Trends Cogn Sci 18, 457-464.
Kruschke, J.K., and Hullinger, R.A. (2010). Evolution of attention in learning. In Computational Models of Conditioning, N.A. Schmajuk, ed. (Cambridge University Press), pp. 10-52.
Lansink, C.S., Goltstein, P.M., Lankelma, J.V., and Pennartz, C.M. (2010). Fast-spiking interneurons of the rat ventral striatum: temporal coordination of activity with principal cells and responsiveness to reward. Eur J Neurosci 32, 494-508.
Le Pelley, M.E., Mitchell, C.J., Beesley, T., George, D.N., and Wills, A.J. (2016). Attention and associative learning in humans: An integrative review. Psychol Bull 142, 1111-1140.
Leong, Y.C., Radulescu, A., Daniel, R., DeWoskin, V., and Niv, Y. (2017). Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. Neuron 93, 451-463.

Lovejoy, L.P., and Krauzlis, R.J. (2010). Inactivation of primate superior colliculus impairs covert selection of signals for perceptual judgments. Nature Neuroscience 13, 261-U153.
Matsumoto, M., Matsumoto, K., Abe, H., and Tanaka, K. (2007). Medial prefrontal cell activity signaling prediction errors of action values. Nat Neurosci 10, 647-656.
Medalla, M., and Barbas, H. (2009). Synapses with inhibitory neurons differentiate anterior cingulate from dorsolateral prefrontal pathways associated with cognitive control. Neuron 61, 609-620.
Mesulam, M.M. (1981). A cortical network for directed attention and unilateral neglect. Annals of neurology 10, 309-325.
Morecraft, R.J., Geula, C., and Mesulam, M.M. (1993). Architecture of connectivity within a cingulo-fronto-parietal neurocognitive network for directed attention. Archives of neurology 50, 279-284.
Morecraft, R.J., Stilwell-Morecraft, K.S., Cipolloni, P.B., Ge, J., McNeal, D.W., and Pandya, D.N. (2012). Cytoarchitecture and cortical connections of the anterior cingulate and adjacent somatomotor fields in the rhesus monkey. Brain research bulletin 87, 457-497.
Niv, Y., Daniel, R., Geana, A., Gershman, S.J., Leong, Y.C., Radulescu, A., and Wilson, R.C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. J Neurosci 35, 8145-8157.
O'Doherty, J.P., Cockburn, J., and Pauli, W.M. (2017). Learning, Reward, and Decision Making. Annual review of psychology 68, 73-100.
Oemisch, M., Westendorff, S., Everling, S., and Womelsdorf, T. (2015). Interareal Spike-Train Correlations of Anterior Cingulate and Dorsal Prefrontal Cortex during Attention Shifts. J Neurosci 35, 13076-13089.
Padoa-Schioppa, C., and Assad, J.A. (2006). Neurons in the orbitofrontal cortex encode economic value. Nature 441, 223-226.
Padoa-Schioppa, C., and Assad, J.A. (2008). The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. Nat Neurosci 11, 95-102.
Paille, V., Fino, E., Du, K., Morera-Herreras, T., Perez, S., Kotaleski, J.H., and Venance, L. (2013). GABAergic circuits control spike-timing-dependent plasticity. J Neurosci 33, 9353-9363.
Pearce, J.M., and Hall, G. (1980). A Model for Pavlovian Learning - Variations in the Effectiveness of Conditioned but Not of Unconditioned Stimuli. Psychological Review 87, 532-552.
Plenz, D., and Kitai, S.T. (1998). Up and down states in striatal medium spiny neurons simultaneously recorded with spontaneous activity in fast-spiking interneurons studied in cortex-striatum-substantia nigra organotypic cultures. J Neurosci 18, 266-283.
Potjans, W., Diesmann, M., and Morrison, A. (2011). An imperfect dopaminergic error signal can drive temporal-difference learning. PLoS computational biology 7, e1001133.
Quilodran, R., Rothe, M., and Procyk, E. (2008). Behavioral shifts and action valuation in the anterior cingulate cortex. Neuron 57, 314-325.
Rasmussen, D., Voelker, A., and Eliasmith, C. (2017). A neural model of hierarchical reinforcement learning. PLoS One 12, e0180234.
Roelfsema, P.R., and van Ooyen, A. (2005). Attention-gated reinforcement learning of internal representations for classification. Neural Comput 17, 2176-2214.
Rombouts, J.O., Bohte, S.M., and Roelfsema, P.R. (2015). How attention can create synaptic tags for the learning of working memories in sequential tasks. PLoS computational biology 11, e1004060.
Saleem, K.S., Kondo, H., and Price, J.L. (2008). Complementary circuits connecting the orbital and medial prefrontal networks with the temporal, insular, and opercular cortex in the macaque monkey. The Journal of comparative neurology 506, 659-693.
Saleem, K.S., Miller, B., and Price, J.L. (2014). Subdivisions and connectional networks of the lateral prefrontal cortex in the macaque monkey. The Journal of comparative neurology 522, 1641-1690.
Santoro, A., Frankland, P.W., and Richards, B.A. (2016). Memory Transformation Enhances Reinforcement Learning in Dynamic Environments. J Neurosci 36, 12228-12242.
Schultz, W. (1998). Predictive reward signal of dopamine neurons. J Neurophysiol 80, 1-27.
Schultz, W. (2016). Dopamine reward prediction-error signalling: a two-component response. Nat Rev Neurosci 17, 183-195.

Schultz, W. (2017). Reward prediction error. Curr Biol 27, R369-R371.
Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. J Neurosci 13, 900-913.
Schultz, W., Tremblay, L., and Hollerman, J.R. (2003). Changes in behavior-related neuronal activity in the striatum during learning. Trends Neurosci 26, 321-328.
Seo, M., Lee, E., and Averbeck, B.B. (2012). Action selection and action value in frontal-striatal circuits. Neuron 74, 947-960.
Serences, J.T. (2008). Value-based modulations in human visual cortex. Neuron 60, 1169-1181.
Seung, H.S. (2003). Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. Neuron 40, 1063-1073.
Shen, C., Ardid, S., Kaping, D., Westendorff, S., Everling, S., and Womelsdorf, T. (2015). Anterior Cingulate Cortex Cells Identify Process-Specific Errors of Attentional Control Prior to Transient Prefrontal-Cingulate Inhibition. Cereb Cortex 25, 2213-2228.
Shenhav, A., Botvinick, M.M., and Cohen, J.D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron 79, 217-240.
Shenhav, A., Cohen, J.D., and Botvinick, M.M. (2016). Dorsal anterior cingulate cortex and the value of control. Nat Neurosci 19, 1286-1291.
Smith, A.C., and Brown, E.N. (2003). Estimating a state-space model from point process observations. Neural Comput 15, 965-991.
Smith, A.C., Frank, L.M., Wirth, S., Yanike, M., Hu, D., Kubota, Y., Graybiel, A.M., Suzuki, W.A., and Brown, E.N. (2004). Dynamic analysis of learning in behavioral experiments. J Neurosci 24, 447-461.
Soltani, A., and Wang, X.J. (2010). Synaptic computation underlying probabilistic inference. Nat Neurosci 13, 112-119.
Suri, R.E., and Schultz, W. (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. Exp Brain Res 121, 350-354.
Sutton, R.S., and Barto, A.G. (1998). Reinforcement Learning: An Introduction (Cambridge, MA: MIT Press).
Takahashi, Y.K., Langdon, A.J., Niv, Y., and Schoenbaum, G. (2016). Temporal Specificity of Reward Prediction Errors Signaled by Putative Dopamine Neurons in Rat VTA Depends on Ventral Striatum. Neuron 91, 182-193.
Urbanczik, R., and Senn, W. (2009). Reinforcement learning in populations of spiking neurons. Nat Neurosci 12, 250-252.
Vogels, T.P., Sprekeler, H., Zenke, F., Clopath, C., and Gerstner, W. (2011). Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. Science 334, 1569-1573.
Voloh, B., Valiante, T.A., Everling, S., and Womelsdorf, T. (2015). Theta-gamma coordination between anterior cingulate and prefrontal cortex indexes correct attention shifts. Proc Natl Acad Sci U S A 112, 8457-8462.
Watabe-Uchida, M., Eshel, N., and Uchida, N. (2017). Neural Circuitry of Reward Prediction Error. Annu Rev Neurosci 40, 373-394.
Watkins, C.J.C.H., and Dayan, P. (1992). Q-Learning. Machine Learning 8, 279-292.
Westendorff, S., Kaping, D., Everling, S., and Womelsdorf, T. (2016). Prefrontal and anterior cingulate cortex neurons encode attentional targets even when they do not apparently bias behavior. J Neurophysiol 116, 796-811.
Wilson, F.A., O'Scalaidhe, S.P., and Goldman-Rakic, P.S. (1994). Functional synergism between putative gamma-aminobutyrate-containing neurons and pyramidal neurons in prefrontal cortex. Proc Natl Acad Sci U S A 91, 4009-4013.
Wilson, R.C., and Niv, Y. (2011). Inferring relevance in a changing world. Frontiers in human neuroscience 5, 189.
Womelsdorf, T., Ardid, S., Everling, S., and Valiante, T.A. (2014). Burst Firing Synchronizes Prefrontal and Anterior Cingulate Cortex during Attentional Control. Curr Biol 24, 2613-2621.

Womelsdorf, T., and Everling, S. (2015). Long-Range Attention Networks: Circuit Motifs Underlying Endogenously Controlled Stimulus Selection. Trends Neurosci 38, 682-700.
Womelsdorf, T., Johnston, K., Vinck, M., and Everling, S. (2010). Theta-activity in anterior cingulate cortex predicts task rules and their adjustments following errors. Proc Natl Acad Sci U S A 107, 5248-5253.
Zenon, A., and Krauzlis, R.J. (2012). Attention deficits without cortical neuronal deficits. Nature 489, 434-437.

Figure legends

Figure 1 | Feature-based reversal learning task and anatomical recording locations.

(A) *Left:* Animals are presented with two black/white stimulus gratings to the left and right of a central fixation point. The stimulus gratings then become colored and start moving in opposite directions. Dimming of the stimuli served as Choice/Go signal. At the time of the dimming of the target stimulus the animals had to indicate the motion direction of the target stimulus by making a corresponding up- or downward saccade in order to receive a liquid reward. Dimming of the target stimulus occurred either before, after or at the same time as the dimming of the distractor stimulus. **(B)** *Left:* Three features characterize each stimulus – color, location, and motion direction. Only the color feature is directly linked to reward outcome. The task is a deterministic reversal learning task, whereby only one color at a time is rewarded. *Right:* This reward contingency switches repeatedly and unannounced in a block-design fashion. **(C)** Average proportion of correct choices relative to the reversal for monkey H (grey) and monkey K (blue). (**D**) Dimension weighted reinforcement learning model with main parameters α, β, ω, and η for feature weighting, selection noise, decay rate and learning rate, respectively. (**E**) Illustration of recording locations relative to stereotaxic zero for monkey H (top) and monkey K (bottom). Neuron locations are collapsed across 5mm coronal slices indicated by the grey bars on the brain on top. Red circles represent neurons that encoded a feature-specific prediction error, grey circles represent neurons that did not. Coronal images displayed are from an average macaque monkey (MRI atlas from (Calabrese et al., 2015).

Figure 2 | Example neurons encoding RPE signals for different feature- and RPE-types and from different areas and monkeys.

For each of six example neurons (**i - vi**), the spike rasters and spike density functions are displayed (for visualization purposes only) for prediction error values of three different magnitudes (trials evenly split into RPE large, RPE medium, RPE small), for the feature value (e.g. color 1) for which an RPE was encoded (top row), and for the feature value for which an RPE was less or not encoded (e.g. color 2, middle row, 'other feature value'). The bottom most row displays the z-transformed R values of the correlation between spike rate and RPE for the two feature values above (solely this last row displays the statistical analyses performed). Black empty (non-significant) or red filled (significant) circles represent z-transformed R values of the correlation between spike rate and RPE for those feature trials for which a RPE was encoded. Grey filled circles with black (non-significant) or red (significant) outlines represent z-transformed R values of the correlation between spike rate and RPE for those feature trials for which a RPE was *not* encoded. Red stars indicate those time bins for which the R-values between the two feature values differed significantly (Z-test, see Methods, eq. 5). Grey transparent bars in *all* plots indicate the time window of RPE encoding. The title above each column of figures indicates the area that neuron was recorded from as well as the type of feature and RPE signal encoded by that neuron. Anatomical images at the top-most additionally illustrate the recording locations. Shaded error bars represent SEM.

Figure 3 | Temporal profile of feature-specific RPE, non-specific RPE and outcome signals.

(**A**) Time courses of feature-specific RPE encoding, non-specific RPE encoding and outcome encoding across all units that encode the given signal combined across both monkeys. For each neuron, *all* time bins for which an RPE was encoded are included.

(**B**) Normalized cumulative sums of the histograms in (A). *Top:* Thick lines represent the mean across both monkeys, while thin continuous lines represent cumulative sums of *monkey H*, and thin dotted lines represent cumulative sums of *monkey K*. All three cumulative sums differed significantly from each other (Kolmogorov-Smirnoff test, Bonferroni-Holm multiple-comparison correction; all $p < .001$). *Bottom:* Magnification of the cumulative sums around the 25% window. Open circles represent the time points at which 25% of the respective signal is encoded. The horizontal bar with three asterisks indicates that all three time points differ significantly from each other (randomization procedure, all $p < .01$).

Figure 4 | Latency comparison of feature-specific RPE encoding across areas.

(A) Time courses of feature-specific RPE encoding in ACC, VS, PFC and CD neurons combined across both monkeys. To enhance visualization of the four histograms lines representing the outlines of each histogram are added.

(B) Normalized cumulative sums of the histograms in (A). *Top:* Thick lines represent the mean across both monkeys, while thin continuous lines represent cumulative sums of *monkey H*, and thin dotted lines represent cumulative sums of *monkey K*. The cumulative sums of all areas except for ACC and CD differed significantly from each other (Kolmogorov-Smirnoff test, Bonferroni-Holm multiple-comparison correction; $p_{ACC\text{-}CD} = .128$, all other $p < .01$). *Bottom:* Magnification of the cumulative sums around the 25% window. Open circles represent the time points at which 25% of feature-specific RPEs is encoded in the four areas. One asterisk indicates $p < .05$; three asterisks indicate $p < .001$ (randomization procedure).

Figure 5 | Prevalence of color-, location- and motion-specific negative and positive RPE encoding.

Shown are proportions of neurons that encode a color-, location-, or motion-specific RPE negative signal either combined across areas **(A)** or split by areas **(B)**. Thick blue lines represent averages across both monkeys. Thin continuous grey lines represent data from *monkey H*, thin dashed grey lines represent data from *monkey K*. An asterisk indicates $p \leq .05$ using a one-sided bootstrap procedure that randomized the feature labels. Dotted lines indicate upper confidence interval. Grey bars indicate chance level proportion at 0.05. **(C)** Color tuning indices for each area computed according to eq. 6. Grey bar represents upper and lower bootstrap confidence interval. An asterisk indicates $p < .05$ by falling outside of the specified confidence interval. **(D-F)** equivalent conventions to **(A-C)** for feature-specific positive RPE encoding.

Figure 6 | Prevalence of color-, location- and motion-specific surprise RPE encoding.

Conventions are equivalent to Figure 7 for feature-specific unsigned RPEs.

Figure 7 | Cell-type classification of RPE units.

(A)-(E) for ACC/PFC units. **(A)** Waveforms of all highly isolated single units recorded, identified as putative interneurons (narrow-spiking, red), putative pyramidal cells (broad-spiking, blue). **(B)** Histogram of the first component of the PCA using peak-to-trough duration and time to repolarization to separate neurons into putative interneurons and putative pyramidal cells. **(C)** Proportion of non-specific RPE encoding neurons identified as narrow- or broad-spiking. **(D)** Proportion of feature-specific RPE encoding neurons identified as narrow- or broad-spiking or unidentified. **(E)** Ratio of narrow to broad spiking neurons identified in the population, for non-specific and feature-specific RPE encoding neurons. Black asterisk indicates $p < 0.05$ (chi-square test). **(F)-(J)** for CD/VS units. **(F)** Waveforms of all highly isolated single units recorded, identified as putative interneurons (red) or putative medium spiny neurons (MSNs, blue), or unidentified (black). **(G)** Histogram of the first component of the PCA using peak width and initial slope of valley decay (ISVD) to separate neurons into putative interneurons and MSNs. Inset shows the distribution of peak width to ISVD across neurons. **(H)** Proportion of non-specific RPE encoding neurons identified as putative interneurons or MSNs. **(I)** Proportion of feature-specific RPE encoding neurons identified as putative interneurons or MSNs or unidentified. **(J)** Ratio of putative interneuron/ MSN in the population, for non-specific and feature-specific RPE encoding neurons.

Figure 8 | Firing rate increases in response to the color onset following low versus high RPE trials.

(A,B) Firing rate changes following color onset in trial n+1 across populations of neurons encoding color-specific positive RPE (*A*) and surprise (unsigned RPE) (*B*). Above zero, rate changes are shown for the 25% of trials with the greatest prediction errors; below zero, rate changes are shown for the 25% of trials with the lowest prediction error, in cyan following preferred color choices and in grey following non-preferred color choices. Thick cyan and grey lines as well as triangles indicate average rate changes. Thin continuous lines show data from *monkey H*, thin dashed lines show data from *monkey K*. (C,D) Mean differences in rate changes following low versus high RPEs for each area areas across color-specific pRPE (*A*), and unsigned RPE (*B*) encoding neurons.

Black asterisks indicate significant differences in rate changes following low versus high RPEs (paired t-test, $p < .05$).

Methods

Electrophysiological recordings

Data was collected from two male rhesus macaques (*Macaca mulatta*). All animal care and experimental protocols were approved by the York University Council on Animal Care and were in accordance with the Canadian Council on Animal Care guidelines. Extra-cellular recordings were made with 1-12 tungsten electrodes (impedance 1.2 - 2.2 MOhm, FHC, Bowdoinham, ME) in anterior cingulate cortex (area 24, ACC), prefrontal cortex (area 46, PFC), caudate nucleus (CD) and ventral striatum (VS) through rectangular recording chambers (20 by 25 mm) implanted over the right hemisphere (see **Figure 1E** and Suppl. Methods for anatomical reconstruction). Electrodes were lowered daily through guide tubes using software controlled precision micro-drives (NAN Instruments Ltd., Israel). Data amplification, filtering, and acquisition were done with a multi-channel acquisition processor (*Neuralynx*). Spiking activity was obtained following a 300 - 8,000 Hz passband filter and further amplification and digitization at 40 kHz sampling rate. Sorting and isolation of single unit activity was performed offline with Plexon Offline Sorter, based on principal component analysis of the spike waveforms. Experiments were performed in a custom-made sound attenuating isolation chamber. Monkeys sat in a custom-made primate chair viewing visual stimuli on a computer monitor (60Hz refresh rate, distance of 58cm). Eye positions were monitored using a video-based eye-tracking system (*EyeLink, SRS Systems*) calibrated prior to each experiment to a 9-point fixation pattern. Eye fixation was controlled within a 1.4-2.0 degree radius window. During the experiments, stimulus presentation, monitored eye positions, and reward delivery were controlled via MonkeyLogic (http://www.brown.edu/Research/monkeylogic/). Liquid reward was delivered by a custom-made, air-compression controlled, mechanical valve system.

Anatomical reconstruction

Recording locations were identified using MRI images obtained following initial chamber placement. During MR scanning, we placed a grid marking the chamber center and peripheral positions as well as a diluted iodine solution inside the chamber for visualization. This allowed the referencing of target regions to the chamber center in the resulting MRI images. Target regions (area 24 – ACC, area 46 – PFC, caudate nucleus, ventral striatum) were identified using the scheme from the Price lab (Saleem et al., 2008; Saleem et al., 2014). The dorsal-ventral positioning of electrodes was estimated daily using the MRI images and audible profiles of spiking activity. The relative coarseness of the MRI images did not allow us to differentiate recording locations in the shell of the nucleus accumbens as opposed to the core of the nucleus accumbens with certainty.

Behavioral paradigm

The monkeys performed a feature-based reversal learning task that required covert spatial attention to one of two stimuli dependent on color-reward associations (**Figure 1A**). These color-reward associations were reversed in an uncued manner between blocks of trials with constant color-reward association (**Figure 1B**). By separating the location of attention from the location of the saccadic response, this task allowed an identification of neural responses to the location of attentional focus independent of neural signals linked to response preparations, during reversal learning. Each trial started with the appearance of a grey central fixation point, which the monkey had to fixate. After 0.5 - 0.9s, two black/white drifting gratings appeared to the left and right of the central fixation point. Following another 0.4s the two stimulus gratings either changed color to black/green and black/red (*monkey K*: black/cyan and black/yellow), or started moving in opposite directions up and down, followed after 0.5 - 0.9s by the onset of the second stimulus feature that had not been presented so far, e.g. if after 0.4s the grating stimuli changed color then after another 0.5 - 0.9s they started moving in opposite directions. After 0.4 - 1s either the red and green stimulus dimmed simultaneously for 0.3s or they dimmed separated by 0.55s, whereby either the red or green stimulus could dim first. The dimming represented the go-cue to make a saccade to one of two response targets displayed above and below the central fixation point. Please note that the monkeys needed to keep central fixation until this dimming event occurred. A saccadic response following the dimming was only rewarded if it was made to the response target that corresponded to the movement direction of the stimulus with the color that was associated with reward in the current block of trials, e.g. if the red stimulus was the currently rewarded target and

was moving upward, a saccade had to be made to the upper response target at the time the red stimulus dimmed. A saccadic response was not rewarded if it was made to the response target that corresponded to the movement direction of the stimulus with the non-reward associated color. A correct response was followed by 0.33ml of water delivered to the monkey's mouth. Across trials of a block the color-reward association remained constant for 30 to a maximum of 100 trials. Performance of 90% rewarded trials (calculated as running average over the last 12 trials) automatically induced a block change. The block change was un-cued, requiring the subject to use the reward outcome they received to learn when the color-reward association was reversed in order to covertly select the stimulus with the rewarded color. In contrast to color, other stimulus features (motion direction or stimulus location) were only randomly related to reward outcome. Saccadic responses had to be initialized within 0.5 s after dimming onset to be considered a choice (rewarded or non-rewarded). All other saccadic responses, e.g. towards the peripheral stimuli, were considered non-choice errors.

We used blocksine gratings with rounded-off edges for the peripheral stimuli, moving within a circular aperture at 0.8 °/s and a spatial frequency of 1.2 (cycles/°) and a radius of 2.0°. Gratings were presented at 5° eccentricity to the left and right of the fixation point.

Data analysis

Analysis was performed with custom MATLAB code (Mathworks, Natick, MA), utilizing functions from the open-source Fieldtrip toolbox (http://www.ru.nl/fcdonders/fieldtrip/). All spike-density functions were smoothed with a Gaussian kernel with a standard deviation of 25ms. Only correct and incorrect choice trials were analyzed, whereby correct choice trials were rewarded trials, while incorrect choice trials were either made to the non-rewarded stimulus or in the incorrect response time window (first vs. second dimming). Fixation breaks, early responses, and no-response trials were not included in any analyses.

Initial neuron selection criteria. Units were only included in any of the following analyses if they i) had a minimum firing rate of 0.5Hz within the feedback epoch (0 - 1.5seconds following feedback onset), ii) prediction errors computed with a reinforcement learning model (see below) could be computed for ≥40 trials, and iii) these minimum of 40 trials could be identified as either occurring *during* learning or *after* learning according to an ideal observer statistics (see below). All trials from blocks that were not learned to criterion were discarded.

Expectation maximization algorithm. To identify at which trial during a block the monkey showed statistically reliable learning we analyzed the monkeys' trial-by-trial choice dynamics using the state–space framework introduced by (Smith and Brown, 2003), and implemented by (Smith et al., 2004). This framework entails a state equation that describes the internal learning process as a hidden Markov or latent process and is updated with each trial. The learning state process estimates the probability of a correct (rewarded) choice in each trial and thus provides the learning curve of subjects. The algorithm estimates learning from the perspective of an ideal observer that takes into account all trial outcomes of subjects' choices in a block of trials to estimate the probability that the outcome in a single trial is correct or incorrect. This probability is then used to calculate the confidence range of observing a correct response. We defined the learning trial as the earliest trial in a block at which the lower confidence bound of the probability for a correct response exceeded the p = 0.5 chance level. The identification of a learning trial allowed to discard blocks that were not learned.

Quantifying prediction errors with reinforcement learning modeling. We quantified the trial-by-trial progression of RPEs during reversal performance using a computational model that combines reinforcement learning (RL) principles with Bayesian tracking of reward probabilities for target features. This hybrid Bayesian-RL model was introduced before (Wilson and Niv, 2012; Niv et al., 2015) to account for behavioral adjustments of choices in a multidimensional visual learning task and was recently validated as a model accounting for feature-based reversal learning in the macaque (Hassani et al., 2017). The model represents the stimuli in terms of their stimulus features (color, motion, location), feature values (color A, color B, downward motion, upward motion, left, right), and the actual combinations of feature values for stimulus 1 and stimulus 2.

The model uses Bayesian inference about which stimulus feature f (color, motion or location) is the likely target feature via $p(f|\mathcal{D}_{1:t})$ to obtain a feature-weighted representation for each stimulus. For tracking target feature probability, we denote the feature dimension as f_d (1: location, 2: direction of motion, 3: color) and for each d, f_d, takes two values 1 and 2. For instance, f_3=1 indicates the first color. We can then calculate the probability for the rewarded stimulus (the target) to have feature d, $p_d = p(d|\mathcal{D}_{1:t}) = \sum_{f_d=1,2} p(f_d|\mathcal{D}_{1:t})$. This defines a feature dimension weight $\phi_d = \frac{p_d^\alpha}{\sum_{d'} p_{d'}^\alpha}$, with exponent α and normalized to yield a sum across dimensions equal to one. The predicted reward value of a feature value is then denoted by W_{f_d} and scaled by the

dimensional weight ϕ_d. The value of the specific stimulus i is given by the sum across all weighted feature values that are part of the stimulus

$$V_i = \sum_d \phi_d W_{f_d} \quad \text{(eq. 1)}$$

The choice of which stimulus is selected on a given trial is implemented with a softmax rule using the Boltzmann function

$$P(C_{t+1} = i) = \frac{\exp(\beta V_{i,t})}{\sum_j \exp(\beta V_{j,t})} \quad \text{(eq. 2)}$$

Following a choice, the stimulus values of the chosen stimulus are updated by a reward prediction error scaled by learning rate η according to:

$$W_{f_d,t+1} = W_{f_d,t} + \eta\left(R_t - V_{i,t}\right) \quad \text{(eq. 3)}$$

Values of the unchosen stimulus feature values were scaled down (decayed) by $(1-\omega)$, similar to previous studies (see Hassani et al., 2017; Niv et al., 2015), according to:

$$W_{f_d,t+1} = (1-\omega)W_{f_d,t} \quad \text{(eq. 4)}$$

In summary, feature values of the chosen stimulus are updated using the RPE (eq. 3) and are separately scaled by a dimensional weight (that may be called attentional weight) calculated using Bayes updating of how the feature dimensions color, motion and location relate to reward outcomes.

We optimized the model by minimizing the negative log likelihood over all trials using up to 20 iterations of the simplex optimization method (matlab function fminsearch) followed by fminunc which constructs derivative information. We used an 80% / 20% (training dataset / test dataset) cross-validation procedure repeated for n=100 times to quantify how well the model predicted the data. Each of the one hundred cross-validations optimized the model parameters on the training dataset. We then quantified the log-likelihood of the independent test dataset given the training datasets optimal parameter values. We validated that the described hybrid Bayes-RL model provides a better fit (lower log-likelihood and lowest Akaike Information Criterion) for the cross-validated test dataset than simpler models that either lacked the Bayesian dimension weighting, or that lacked the decay of nonchosen stimulus features (for a detailed evaluation of different models, see also (Hassani et al., 2017)).

Both monkeys choice data were fit well by the Bayes-RL with log likelihoods for *monkey H* and *monkey K* of 0.47 and 0.52, respectively). The model parameters best explaining the data for *monkey H*/*K* had a similar pattern with η (learning rate) = 0.22/0.25, β (selection noise) =

3.55/2.79, ϕ (dimension weighting of feature representation) = 0.68/0.98 and ω (value decay for nonchosen feature) = 0.92/0.68. These results resonate well with previous studies using a similar model architecture (Wilson and Niv, 2012; Niv et al., 2015; Hassani et al., 2017; Leong et al., 2017).

Identification of prediction error encoding neurons. To identify RPE encoding neurons, we correlated each neuron's firing rate time-resolved with RPEs obtained from the RL model. Each correlation analysis required a minimum of 15 trials. We correlated firing rate with positive RPEs in correct choice trials and with negative RPEs in incorrect choice trials. To identify neurons that encoded an unsigned RPE, we used partial correlation analysis to correlate firing rates with the absolute RPE in correct and incorrect choice trials while partializing out the sign of the RPE (by including a co-variate of +/-1 for correct/incorrect trials respectively). The analysis time ranged from -500 to 1500ms after the outcome event; time windows spanned 200ms and were shifted by 25ms. For a neuron to be considered to encode a non-specific positive, negative, or unsigned RPE signal, it had to significantly positively correlate its firing rate with a positive, or unsigned RPE, respectively (Spearman correlation, $p < 0.05$), for a minimum of four consecutive time bins following the outcome event, while not correlating positively in more than two consecutive time bins before the outcome event. For a neuron to be considered to encode a negative RPE signal, it had to significantly negatively correlate its firing rate with a negative RPE for a minimum of four consecutive time bins following the outcome event (Spearman correlation, $p < 0.05$), while not correlating negatively in more than two consecutive time bins before the outcome event.

To identify neurons that encoded a feature-specific RPE signal, trials were split into the features of interest prior to the correlation analysis (color, location and motion direction). The principle for identifying positive, negative and unsigned feature-specific RPE neurons was the same as for non-specific RPE signals with additional criteria described in the following. For instance, for a neuron to be considered to encode a color-specific RPE signal, it had to significantly encode a RPE signal (as described above) in minimally four consecutive time bins for trials in which e.g. color 1 was chosen, while either not encoding or encoding significantly less a RPE signal for trials in which color 2 was chosen. Significant differences between R values (Spearman correlation) for the two trial types were computed by z-transforming R values and comparing them using a z-test:

$$Z_{observed} = \frac{z_1 - z_2}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}} \quad \text{(eq. 5)}$$

where z_1 and z_2 are the z-transformed R-values for the correlation with feature value 1 and feature value 2, respectively. When $Z_{observed}$ exceeded |1.96| ($p<0.05$), R values were considered significantly different for a given time bin. In a minimum of four consecutive bins, R values from correlations with two different feature values (e.g. color 1 chosen or color 2 chosen) had to significantly differ, while a RPE had to be encoded for at least one of the two feature values according to the same criteria as for non-specific RPE signals. The method of identification was the same for identifying location and motion-specific RPE signals, with the exception of splitting trials according to chosen location or chosen motion direction, respectively. We determined for each neuron the duration in which it encoded a RPE signal as the first span of four or more consecutive significant time bins after the feedback event.

Time courses of prediction error and trial outcome signal encoding. To compare time courses of RPE signals, as well as trial outcome signals, we determined for each neuron the time window (minimum 4 consecutive bins) in which it encoded a RPE/trial outcome signal significantly (if a neuron encoded an RPE/trial outcome signal over longer time spans with time bins in between that were not significant, only the first time window of consecutive significant time bins was considered for this analysis). Across neurons, we therefore obtained distributions of time bins in which RPE/trial outcome signals were encoded, and we then tested these distributions for differences in their cumulative sums (Kolmogorov-Smirnoff test, Bonferroni-Holm multiple comparison correction, $\alpha = 0.05$). As an additional measure of latency, we tested whether the time point at which 25% of RPE/trial outcome signals were encoded (the time point when the respective cumulative sum reaches 25%) differed using a randomization procedure ($\alpha = 0.05$). The analysis procedure was equivalent when comparing the latencies of feature-specific RPE encoding between areas.

Comparing the prevalence of prediction error encoding. We used a bootstrap procedure to determine whether any feature-specific RPEs (color, location, or motion direction) were encoded more prevalently than would be expected based on the distribution across all feature-specific RPEs independent of their specificity (n=10,000). This bootstrap procedure was computed across all units encoding a specifically signed or unsigned RPE, initially independent of area recorded, and in a second step separately for each area. Significance was determined based on the actual

proportion of color-, location-, or motion-specific RPEs falling inside or outside of the one-sided confidence interval of the bootstrap procedure. To compare the ratio of color-specific RPE encoding versus location- or motion-specific RPE encoding between areas, we computed a color tuning index for each area as follows:

$$I_{col} = \frac{P_{col} - (P_{loc} + P_{mot})/2}{P_{col} + P_{loc} + P_{mot}} \quad \text{(eq. 6)}$$

whereby I_{col} refers to the color tuning index, P_{col}, P_{loc}, and P_{mot} refer to the proportions of color-, location-, and motion-specific RPE units, respectively. We then compared color tuning indices across areas by computing a confidence interval (bootstrap procedure, n=10,000) around color-tuning indices that were computed with randomized area labels. An area was considered to have a significantly greater or smaller color tuning index than the other areas if it fell outside of the confidence interval.

Cell-type specificity of RPE encoding neurons. For the set of highly isolated neurons (*monkey H*: n = 428, *monkey K*: 398), we aligned, normalized, and averaged all action potentials (Ardid et al., 2015). To distinguish putative interneurons (narrow-spiking) and putative pyramidal cells (broad-spiking) in PFC and ACC, we analyzed the peak-to-trough duration and the time for repolarization for each neuron (Oemisch et al., 2015). The time for repolarization was defined as the time at which the waveform amplitude decayed 15% from its peak value. We computed the principal component analysis (PCA) and used the first component because it allowed for better discrimination between narrow- and broad-spiking cells, compared to any of the two measures alone (Hartigan dip test, $p < 0.0005$). In addition, a comparison of Akaike's and Bayesian was used to confirm that a two-Gaussian model fit the data better than a one-Gaussian model. To distinguish putative interneurons and putative medium-spiny neurons (MSNs) in CD and VS, we analyzed the peak width (at half maximum) and *I*nitial *S*lope of *V*alley *D*ecay (ISVD, Berke, 2008; Lansink et al., 2010), as they provided a better waveform discrimination than e.g. peak-to-trough duration. The ISVD was computed as follows:

$$ISVD = 100 * \frac{(V_T - V_{0.26})}{A_{PT}}$$

where V_T is the most negative value (trough) of the spike waveform, $V_{0.26}$ is the voltage at 0.26 ms after V_T, and A_{PT} is the peak-to-trough amplitude (Lansink et al., 2010). Although we could not discard unimodality for the first PCA component (or for either of the single measures, Hartigan dip test, $p > 0.05$), Akaike's and Bayesian information criteria confirmed that a two-Gaussian

model fit the data better than a one-Gaussian model. For frontal and cingulate units, we then used the two-Gaussian model and divided neurons into two groups of narrow and broad spiking units. For striatal units, because we could not discard unimodality for the first PCA component, we used the two-Gaussian model and defined two cutoffs that divided neurons into three groups. The first cutoff was defined as the point at which the likelihood of a narrow-spiking/putative interneuron was 3 times larger than the likelihood of a broad-spiking/putative principal cell, and vice versa for the second cutoff. We reliably classified PFC/ACC neurons (n = 485) as either putative pyramidal cells (broad spiking, n = 344, *monkey H*: 203, *monkey K*: 183) or putative interneurons (narrow-spiking, n = 141, *monkey H*: 78, *monkey K*: 49). Therefore, in *monkey H* 72% of neurons in ACC/PFC were identified as putative pyramidal cells while 28% of neurons were identified as putative interneurons. In *monkey K*, 82% of neurons were identified as putative pyramidal cells and 17% as putative interneurons. We classified 96% of striatal neurons (n = 277) as either putative MSNs (broad spiking, n = 198, *monkey H*: 96, *monkey K*: 113) or putative interneurons (narrow-spiking, n = 79, *monkey H*: 35, *monkey K*: 36), while n = 26 (*monkey H*: 8, *monkey K*: 11) neurons fell in between the criteria and could not be reliably classified. Therefore, in *monkey H* 73% of neurons in CD/VS were identified as putative MSNs while 27% of neurons were identified as putative interneurons. In *monkey K*, 77% of neurons were identified as putative MSNs and 23% as putative interneurons. For striatal units, we additionally verified our classification by comparing the firing rates between neurons classified as MSNs and those classified as interneurons. Striatal interneurons tend to be fast-spiking interneurons and should have a higher firing rate than the relatively low-firing MSNs (Berke et al., 2004; Berke, 2008). Indeed, in both monkeys, neurons classified as interneurons had a higher mean firing rate (*monkey H*: 4.96±1.1 Hz, *monkey K*: 4.77 ± 2.62 Hz) than neurons classified as MSNs (*monkey H*: 1.70 ± 0.38 Hz, *monkey K*: 1.61 ± 0.26 Hz), and this was statistically reliable in both monkeys (t-test, *monkey H*: $p < 0.001$, *monkey K*: $p = 0.039$). For the analysis of narrow versus broad spiking feature-specific versus non-specific RPE units we combined data from both monkeys because of relatively low neuron numbers. Proportions of narrow versus broad spiking units between non-specific and feature-specific RPE neurons were compared using chi-square statistics.

Stimulus selection following low and high prediction errors. We tested how neurons that encoded a color-specific prediction error changed their firing rate during color selection in trials following low versus high prediction errors (**Figure 7**). To do so, we identified for each color-

specific RPE neuron the 25% of trials with the greatest prediction errors (from the model) and those 25% with the lowest prediction errors. These trials were then split into whether the choice was made to the color for which an RPE was encoded (preferred color) and those for which a choice was made to the other color (non-preferred color). For each trial n we found trial n+1 and computed the change in firing rate from the 400ms prior to stimulus color onset to the 100-700ms following stimulus color onset (post-color – pre-color / post-color + pre-color). We thus computed for each neuron the average rate change following low RPE trials in which the preferred color was chosen, following low RPE trials in which the non-preferred color was chosen, following high RPE trials in which the preferred color was chosen, and following high RPE trials in which the non-preferred color was chosen. Across the populations of color-specific positive RPE, negative RPE, or unsigned RPE encoding neurons, we then compared the change in firing rate at stimulus selection following high versus low RPE trials for the preferred color and for the non-preferred color choice trials using paired t-tests. In a second step, we split neuron populations based on their respective recording locations and performed the equivalent analysis.

Task variables encoded in the outcome epoch. To characterize neural responses in the outcome epoch, we adapted analyses from Padoa-Schioppa and colleagues (Cai and Padoa-Schioppa, 2014; Padoa-Schioppa and Assad, 2006, 2008). We tested whether neurons encoded any of twelve variables at the time of reward onset/omission. These twelve variables included the three stimulus features (color, location, motion) i) selected in the current choice independent of choice outcome (correct and error) (**chosen color, chosen location, chosen motion**) (Genovesio et al., 2014), ii) selected in the previous choice (trial n-1) independent of choice outcome (correct and error) (**previous chosen color, previous chosen location, previous chosen motion**) (Donahue and Lee, 2015; Genovesio et al., 2014), iii) of the target independent of choice (correct and error) (**target color, target location, target motion**) (Westendorff et al., 2016), in addition to the variables **outcome** (correct and error), **previous outcome** (correct and error) (Donahue and Lee, 2015) and **learning progress** (correct trials *during* learning versus *after* learning as obtained from the EM algorithm described above). To estimate the correlation between variables, we computed the correlation coefficient between any two trial vectors of the variables per recording session and then computed the average absolute correlation coefficient across recording sessions (the average correlation coefficient now varies between 0 and 1). The correlation matrix is shown in **Suppl. Fig. S3C**. To identify whether any neuron encodes any one variable, we performed

independent linear regressions for each neuron on each variable. A neuron's firing rate was averaged in the 0.1 - 0.7 seconds after reward onset/omission and was considered to significantly encode a variable at $p \leq 0.05$. In general, a neuron's response could be explained by multiple variables, which is likely because variables are correlated with each other, a situation referred to as multi-collinearity. We therefore adapted the "best-subset" method as a method of variable selection used in the case of multi-linear regressions (Dunn and Clark, 1987; Glantz and Slinker, 2001; Padoa-Schioppa and Assad, 2006).

Best-subset method. We computed for each subset of d variables the total number of neural responses explained by that subset and determined which subset explained the maximum number of responses. This was repeated for d=1, 2, 3.. variables per subset. We determined the number of variables necessary to characterize the population when 85% of the maximum number of responses explained was reached. The best-subset method assumes that each neuron only encodes a single variable. We therefore tested for second-order encoding to determine the proportion of neurons that encoded more than one variable (Padoa-Schioppa and Assad, 2006).

Second order encoding. We found for each neuron the best-fit variable and its corresponding R^2 value. To determine whether adding an additional variable to the regression led to a significantly higher R^2 value, we computed:

$$F_{X,Y}=\frac{(n-3)*(R^2_{XY}-R^2_X)}{(1-R^2_{XY})}$$

where R^2_X is from the original linear regression on X only, R^2_{XY} is from the bilinear regression on X and Y and n is the number of trials. $F_{X,Y}$ is computed for each of the eleven possible second variables and the maximum F is found. If the corresponding p-value for the maximum F value is ≤ 0.01, we consider the neuron to significantly encode the second variable (Padoa-Schioppa and Assad, 2006). 31% of neurons significantly encoded a second task variable, which is more than expected by chance (binomial test, $p < .0001$). The major variables that were multiplexed at the second order were previous trial outcome (17.8%) and learning progress (26.7%), with both of these more often encoded at the second order than expected based on an equal distribution across all twelve variables (binomial test, $p < .001$).

Figure 1

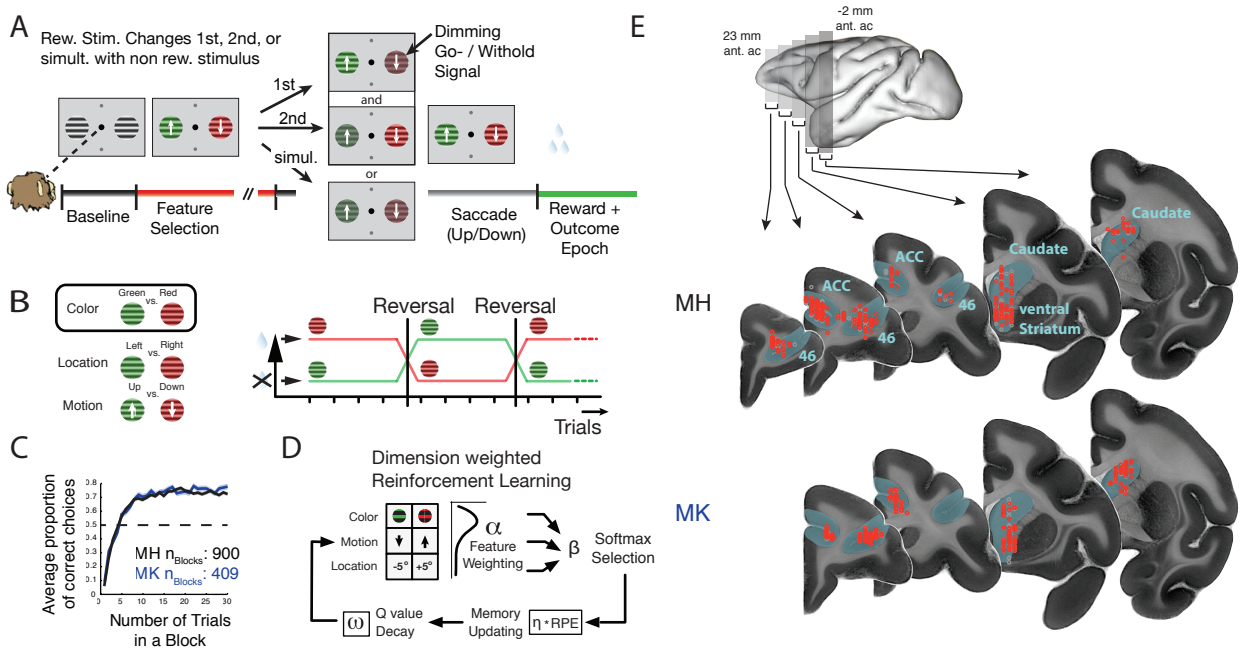

A
Rew. Stim. Changes 1st, 2nd, or simult. with non rew. stimulus
1st
2nd
simul.
and
or
Dimming Go- / Withold Signal
Baseline
Feature Selection
Saccade (Up/Down)
Reward + Outcome Epoch
B
Color
Green vs. Red
Location
Left vs. Right
Motion
Up vs. Down
Reversal
Reversal
Trials
C
Average proportion of correct choices
MH n Blocks: 900
MK n Blocks: 409
Number of Trials in a Block
D
Dimension weighted Reinforcement Learning
Color
Motion
Location
-5°
+5°
α
Feature Weighting
β
Softmax Selection
ω
Q value Decay
Memory Updating
η * RPE
E
-2 mm ant. ac
23 mm ant. ac
MH
MK
ACC
Caudate
ventral Striatum
46

Figure 2

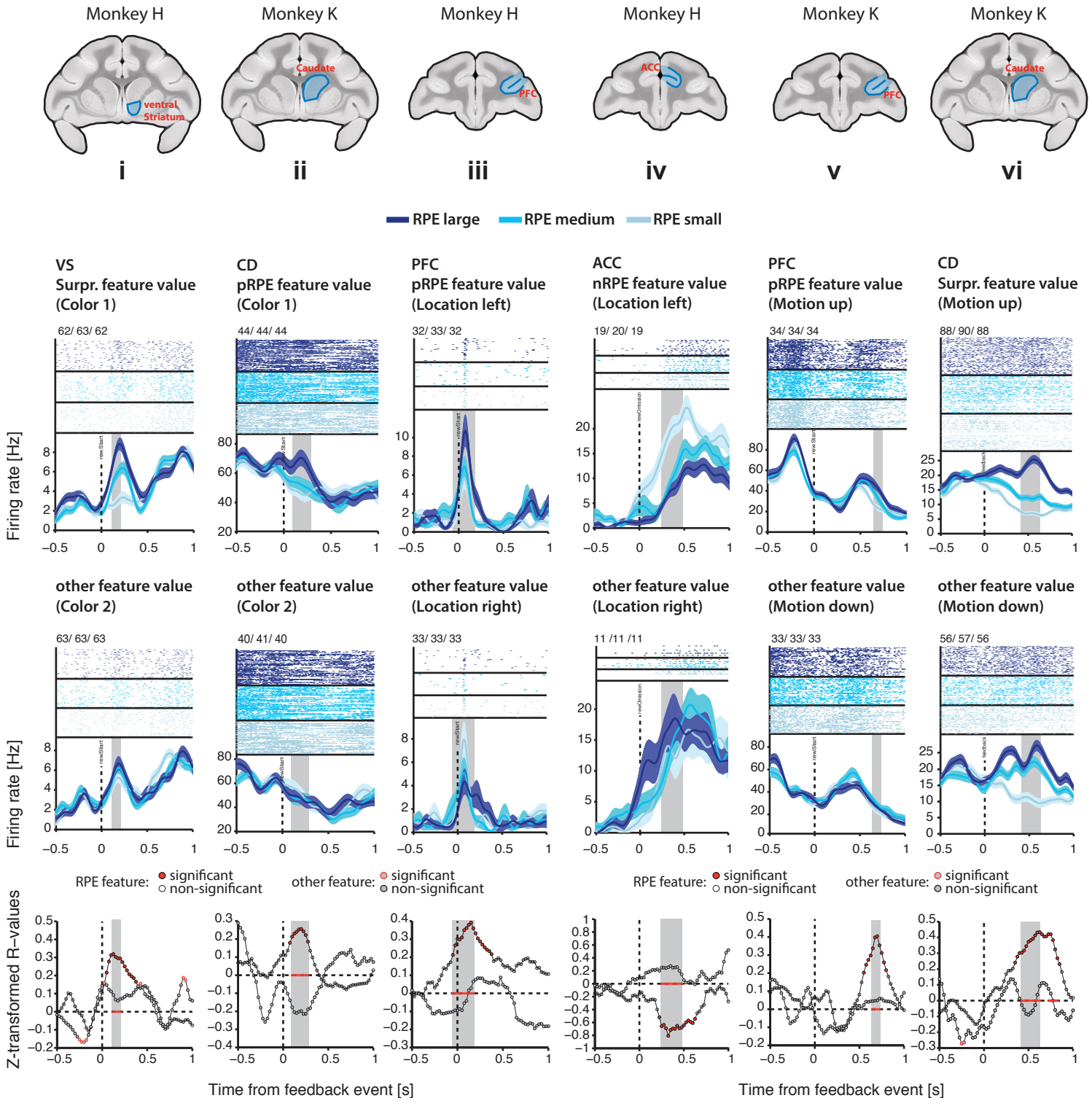

Monkey H
Monkey K
Monkey H
Monkey H
Monkey K
Monkey K
ventral Striatum
Caudate
PFC
ACC
PFC
Caudate
i
ii
iii
iv
v
vi
RPE large
RPE medium
RPE small
VS
Surpr. feature value (Color 1)
CD
pRPE feature value (Color 1)
PFC
pRPE feature value (Location left)
ACC
nRPE feature value (Location left)
PFC
pRPE feature value (Motion up)
CD
Surpr. feature value (Motion up)
62/ 63/ 62
44/ 44/ 44
32/ 33/ 32
19/ 20/ 19
34/ 34/ 34
88/ 90/ 88
Firing rate [Hz]
other feature value (Color 2)
other feature value (Color 2)
other feature value (Location right)
other feature value (Location right)
other feature value (Motion down)
other feature value (Motion down)
63/ 63/ 63
40/ 41/ 40
33/ 33/ 33
11 /11 /11
33/ 33/ 33
56/ 57/ 56
Firing rate [Hz]
RPE feature: significant non-significant
other feature: significant non-significant
RPE feature: significant non-significant
other feature: significant non-significant
Z-transformed R–values
Time from feedback event [s]
Time from feedback event [s]

Figure 3

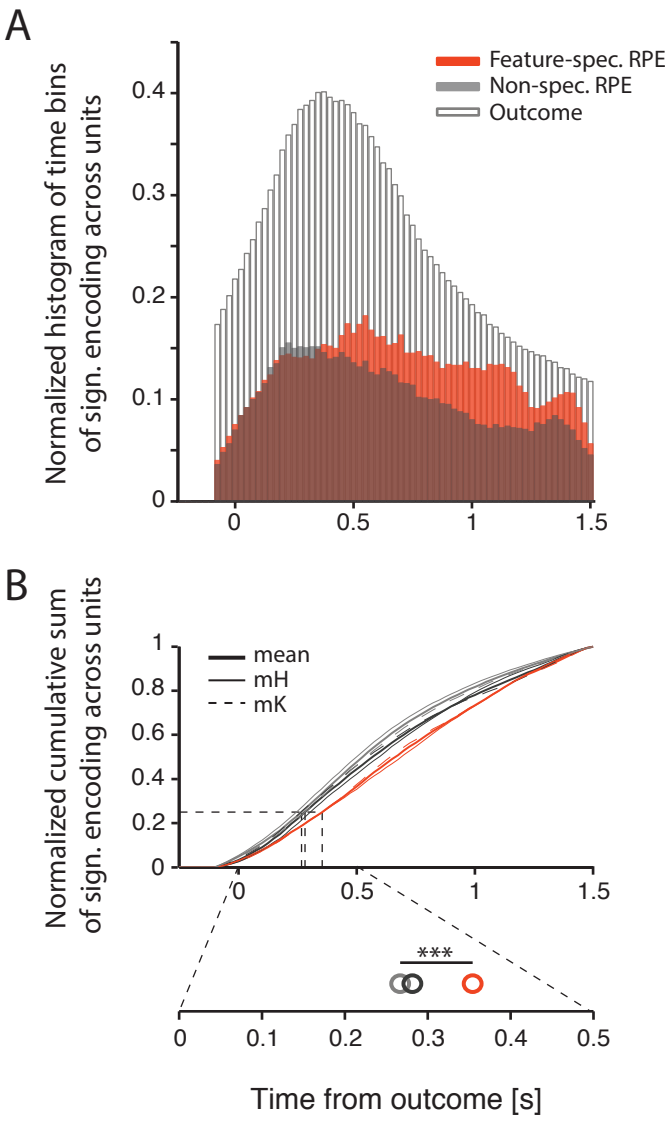

A
Feature-spec. RPE
Non-spec. RPE
Outcome
Normalized histogram of time bins of sign. encoding across units
0.4
0.3
0.2
0.1
0
0
0.5
1
1.5
B
Normalized cumulative sum of sign. encoding across units
mean
mH
mK
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
***
0
0.1
0.2
0.3
0.4
0.5
Time from outcome [s]

Figure 4

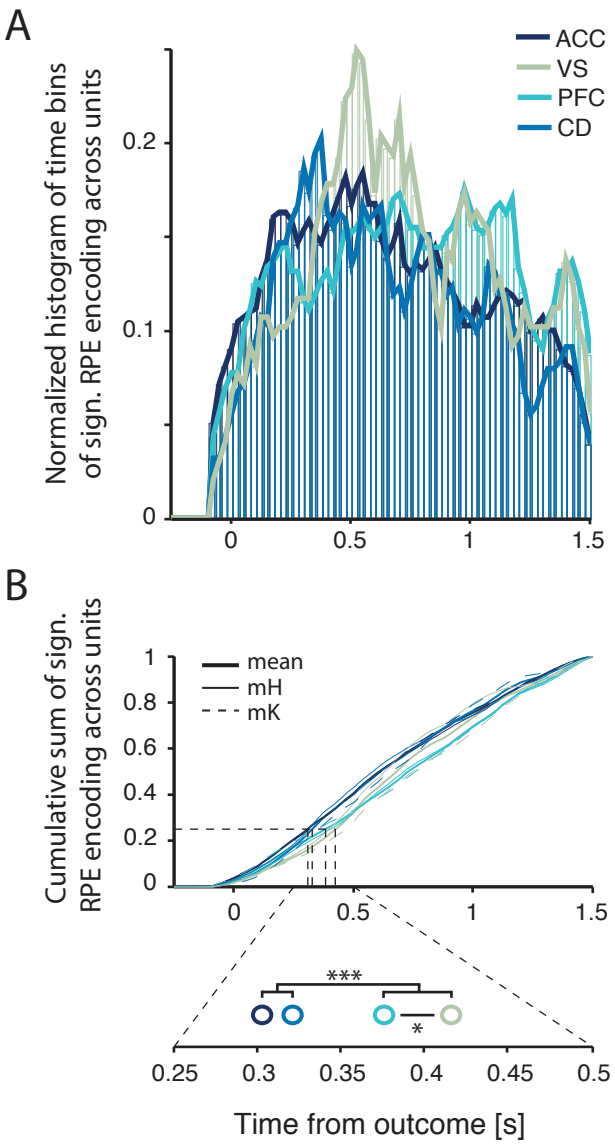

A
ACC
VS
PFC
CD
Normalized histogram of time bins of sign. RPE encoding across units
0.2
0.1
0
0
0.5
1
1.5
B
Cumulative sum of sign. RPE encoding across units
mean
mH
mK
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
***
*
0.25
0.3
0.35
0.4
0.45
0.5
Time from outcome [s]

Figure 5

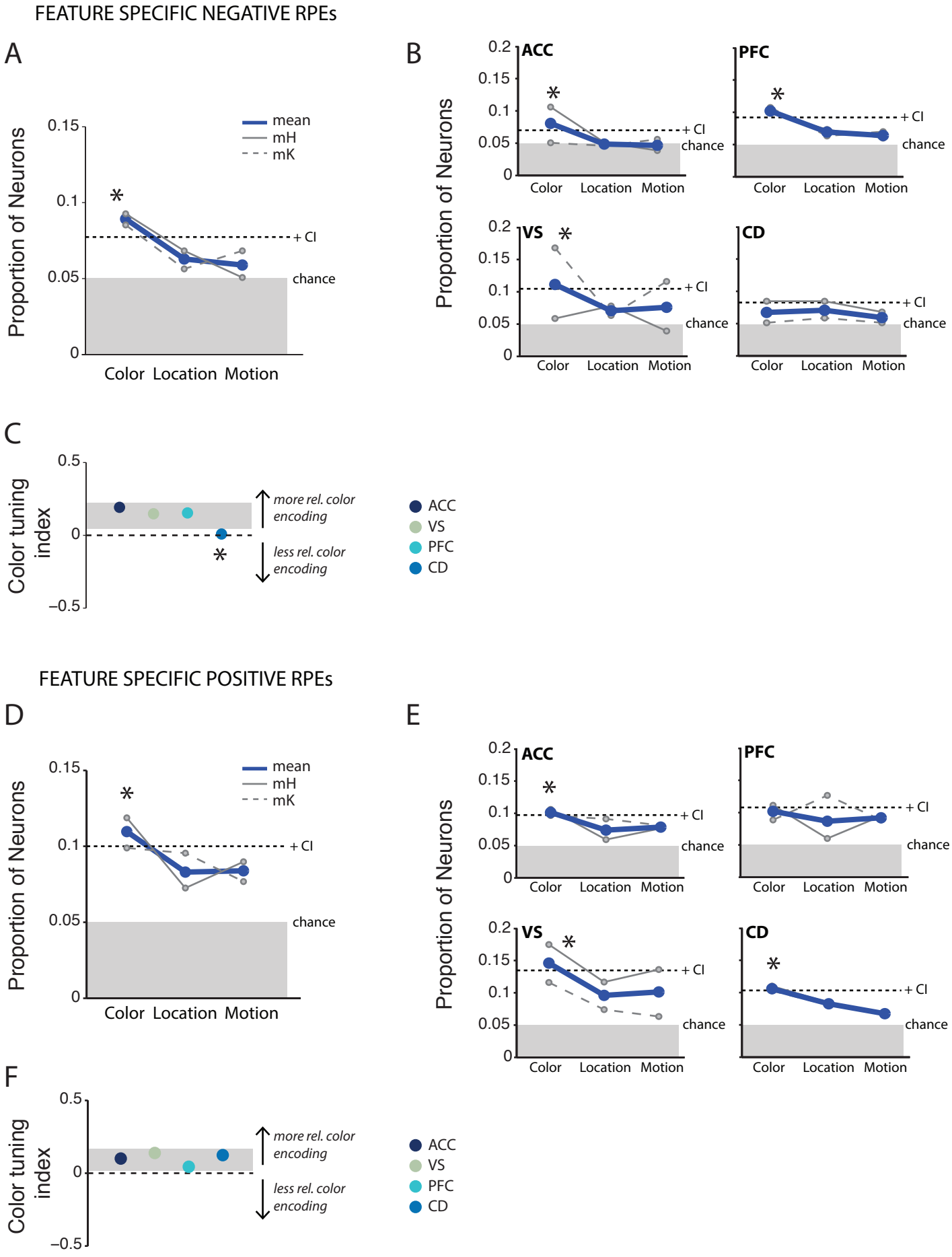

FEATURE SPECIFIC NEGATIVE RPEs
A
B
Proportion of Neurons
mean
mH
mK
+ CI
chance
Color Location Motion
ACC
PFC
VS
CD
C
Color tuning index
more rel. color encoding
less rel. color encoding
ACC
VS
PFC
CD
FEATURE SPECIFIC POSITIVE RPEs
D
E
F

Figure 6

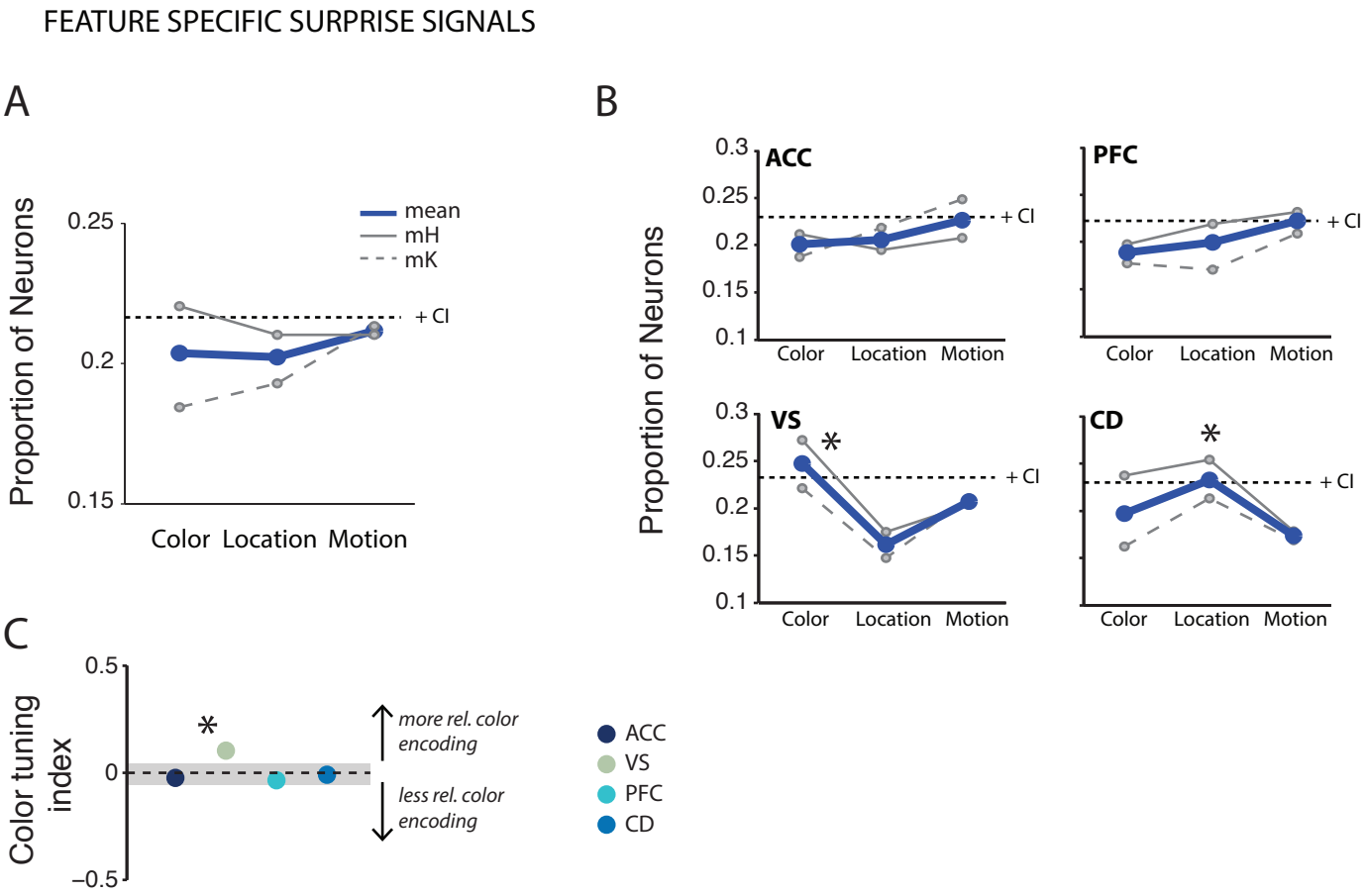

FEATURE SPECIFIC SURPRISE SIGNALS
A
Proportion of Neurons
0.25
0.2
0.15
mean
mH
mK
+ CI
Color Location Motion
B
Proportion of Neurons
ACC
PFC
VS
CD
0.3
0.25
0.2
0.15
0.1
+ CI
Color Location Motion
C
Color tuning index
0.5
0
–0.5
more rel. color encoding
less rel. color encoding
ACC
VS
PFC
CD

Figure 7

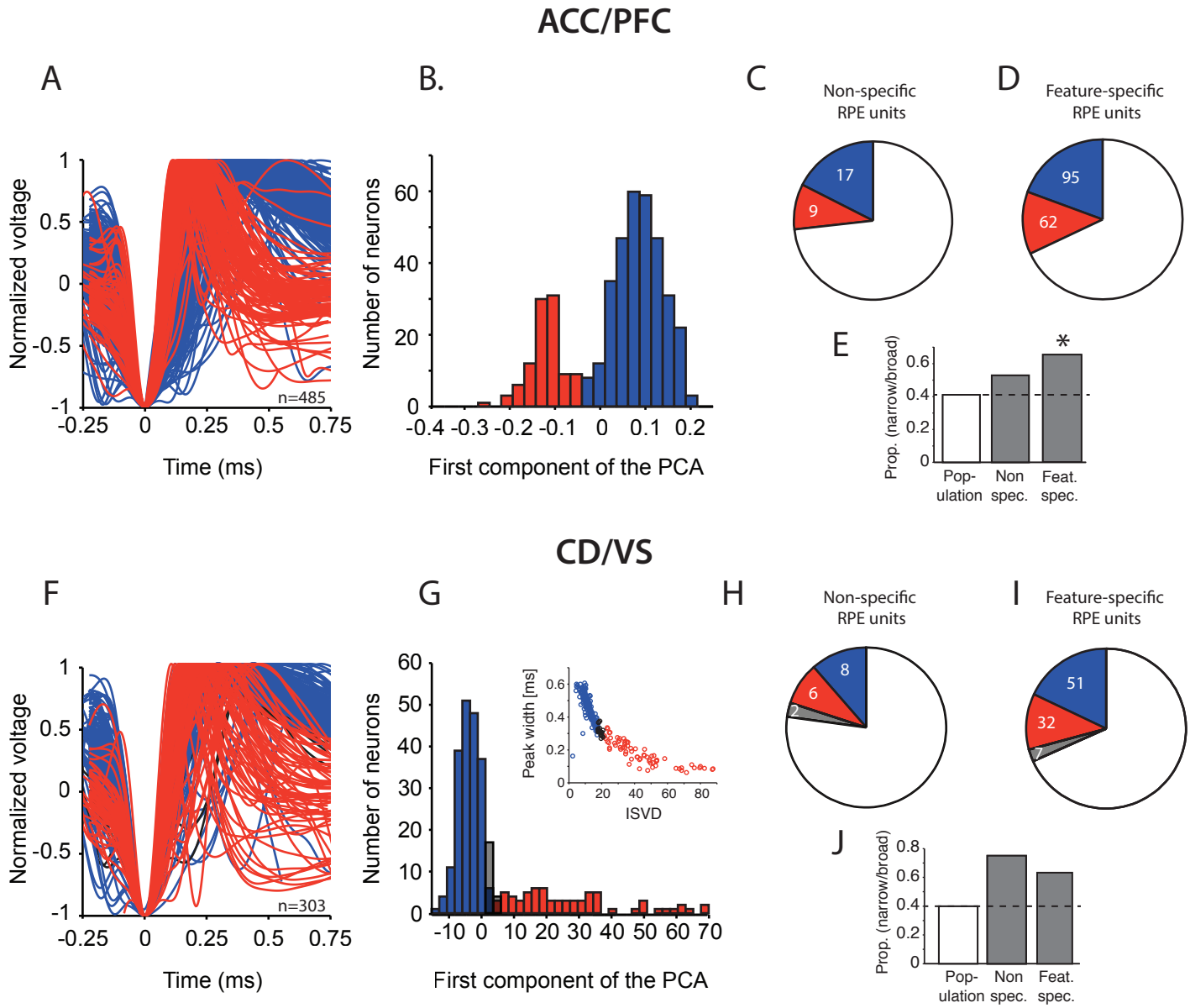
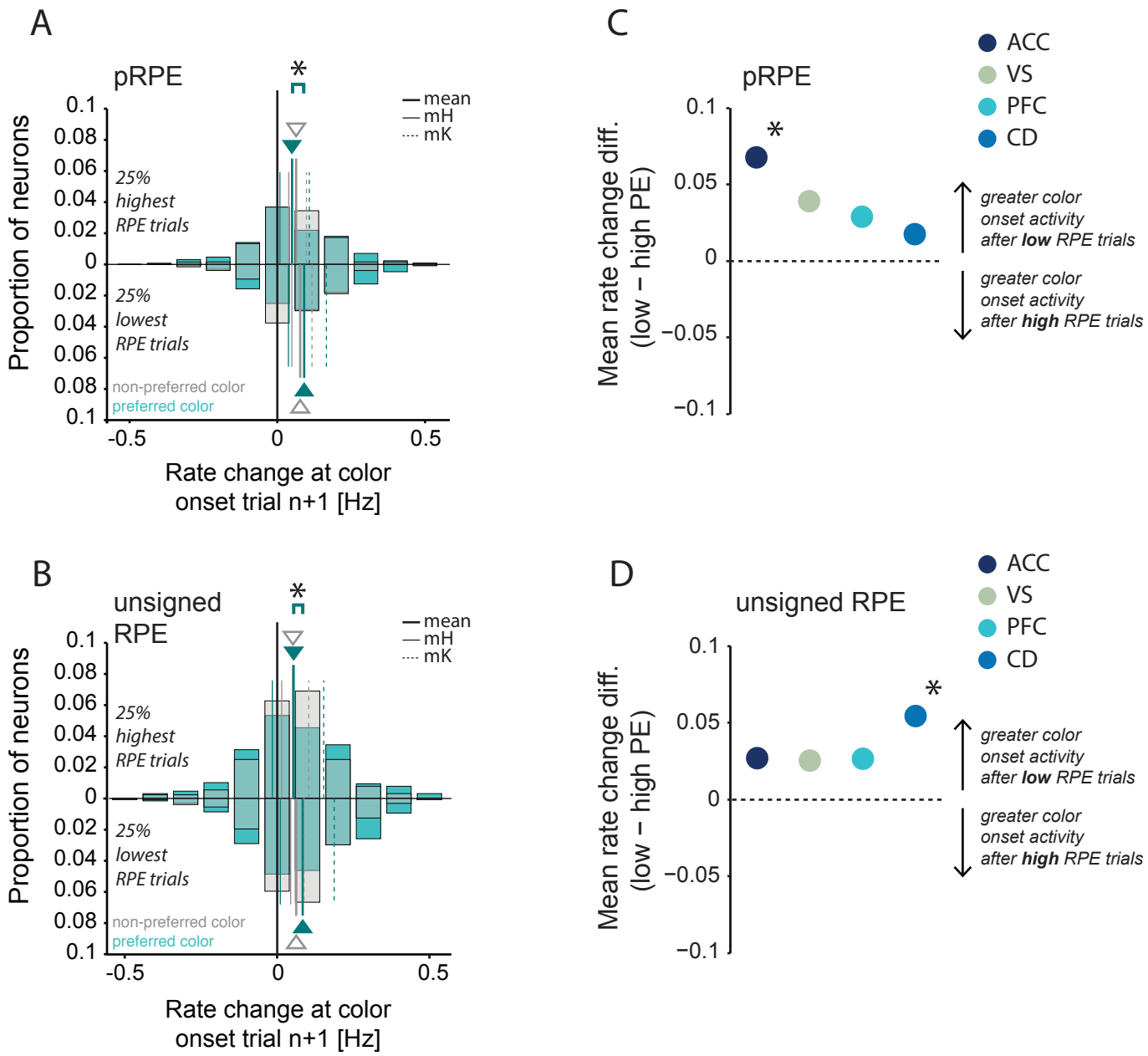

ACC/PFC
A
B.
C
D
Non-specific RPE units
Feature-specific RPE units
17
9
95
62
E
Normalized voltage
n=485
Time (ms)
Number of neurons
First component of the PCA
Prop. (narrow/broad)
Pop-ulation
Non spec.
Feat. spec.
*
CD/VS
F
G
H
I
J
n=303
Peak width [ms]
ISVD
8
6
2
51
32
7

Figure 8


A
pRPE
*
mean
mH
mK
Proportion of neurons
0.1
0.08
0.06
0.04
0.02
0
0.02
0.04
0.06
0.08
0.1
25%
highest
RPE trials
25%
lowest
RPE trials
non-preferred color
preferred color
-0.5
0
0.5
Rate change at color
onset trial n+1 [Hz]
B
unsigned
RPE
*
mean
mH
mK
Proportion of neurons
25%
highest
RPE trials
25%
lowest
RPE trials
non-preferred color
preferred color
-0.5
0
0.5
Rate change at color
onset trial n+1 [Hz]
C
pRPE
ACC
VS
PFC
CD
Mean rate change diff.
(low – high PE)
0.1
0.05
0
−0.05
−0.1
*
greater color
onset activity
after low RPE trials
greater color
onset activity
after high RPE trials
D
unsigned RPE
ACC
VS
PFC
CD
Mean rate change diff.
(low – high PE)
0.1
0.05
0
−0.05
−0.1
*
greater color
onset activity
after low RPE trials
greater color
onset activity
after high RPE trials