Published in final edited form as:
Neuron. 2013 October 2; 80(1): 223–234. doi:10.1016/j.neuron.2013.07.040.

Cortical Signals for Rewarded Actions and Strategic Exploration

Christopher H. Donahue[1,†], **Hyojung Seo**[1,†], and **Daeyeol Lee**[1,2,3,*]

[1]Department of Neurobiology, Yale University School of Medicine, New Haven, CT 06510

[2]Kavlie Institute for Neuroscience, Yale University School of Medicine, New Haven, CT 06510

[3]Department of Psychology, Yale University, New Haven, CT 06520

SUMMARY

In stable environments, decision makers can exploit their previously learned strategies for optimal outcomes, while exploration might lead to better options in unstable environments. Here, to investigate the cortical contributions to exploratory behavior, we analyzed singleneuron activity recorded from 4 different cortical areas of monkeys performing a matching pennies task and a visual search task, which encouraged and discouraged exploration, respectively. We found that neurons in multiple regions in the frontal and parietal cortex tended to encode signals related to previously rewarded actions more reliably than unrewarded actions. In addition, signals for rewarded choices in the supplementary eye field were attenuated during the visual search task, and were correlated with the tendency to switch choices during the matching pennies task. These results suggest that the supplementary eye field might play a unique role in encouraging animals to explore alternative decision-making strategies.

INTRODUCTION

The conflict between exploration and exploitation is common in everyday life. For example, when you visit your favorite restaurant, do you choose to order the same meal that has brought you pleasure in the past, or do you explore the menu for a new dish that could perhaps become your new favorite? In uncertain real-world environments, animals constantly face a choice between exploiting familiar but potentially less valuable options, and exploring unknown options that have the potential to lead to greater rewards. Although the explorationexploitation dilemma plays a central role in reinforcement learning (Sutton and Barto, 1998), how this is resolved in the brain remains poorly understood. While animals can learn to exploit actions that led to reward by simply strengthening associations between stimuli and rewarded actions, they must be able to explore alternative actions in a dynamically changing environment. Thus, uncertainty in the animal's environment might drive exploratory behaviors, and neuromodulators such as norepinephrine may play a role in promoting exploration (Aston-Jones and Cohen, 2005; Yu and Dayan, 2005). In addition, the activity in the fronto-polar cortex is correlated with exploratory choice behaviors (Daw et al., 2006). Nevertheless, relatively little is known about whether and how other cortical regions might play a role in promoting exploration versus exploitation. In addition,

[*] correspondence to: daeyeol.lee@yale.edu.

[†] These authors contributed equally to the study.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

exploration might be necessary to identify appropriate highorder strategies, such as specific sequences of actions (Averbeck et al., 2006), but the neural substrates of such strategic explorations have not been studied.

In the present study, we tested how different cortical regions might contribute to strategic exploration by training rhesus monkeys to perform two tasks which required different levels of exploration and exploitation. First, animals performed a simple visual search task in which correct target locations were explicitly cued. In this task, the animals were required to make choices according to a fixed rule which never changed throughout the entire experiment. In the second task, known as matching pennies, the animals were encouraged to choose randomly, independently from their previous choices and outcomes, since the computer opponent penalized any stereotypical choice sequences (Barraclough et al., 2004; Lee et al., 2004; Seo and Lee, 2007; Seo et al., 2009). Thus, this task discouraged the animals from adopting any deterministic strategies, such as the so-called win-stay strategy, and kept them in a constant state of exploration. We tested whether the strength of signals related to the animal’s recent action varied according to their outcomes and task demands. We tested this for 4 cortical areas, supplementary eye field (SEF), dorsolateral prefrontal cortex (DLPFC), dorsal anterior cingulate cortex (ACC), and lateral intraparietal area (LIP), all of which contain signals related to the animal’s previous choices and their outcomes (Barraclough et al., 2004; Seo and Lee, 2007; Seo et al., 2009; Seo and Lee, 2009). We found that signals related to past actions were encoded more robustly following rewarded than non-rewarded trials in all cortical areas except for ACC. In addition, the previous choice signals in the SEF displayed two important features, in that they were correlated with the animal’s tendency to switch during the matching pennies task, and were also attenuated during the visual search task. Taken together, these results suggest that SEF might play a unique role in promoting exploratory behavior.

RESULTS

Reinforcement Learning during Matching Pennies Task

Six rhesus monkeys were trained to play a competitive game known as matching pennies (Figure 1A). During this task, the animals were required to make a choice between two identical targets, and were rewarded if they chose the same target as a simulated computer opponent. The computer was programmed to exploit statistical biases in the animals’ behavior, such that they had to choose each target with an equal probability and independently across successive trials in order to maximize rewards (see Experimental Procedures).

As reported previously, the choice behavior of the monkeys during the matching pennies task was highly stochastic (Lee et al., 2004). The optimal strategy, known as Nash equilibrium (Nash, 1950), for the matching pennies task used in our study is to choose each target with 0.5 probability. The percentage of trials in which the animals made rightward choices ranged from 48.7% to 51.0% across animals (Table 1), indicating that the matching pennies task prevented the animals from developing a strong spatial bias. The computer opponent chose one of the two targets randomly when it failed to detect any significant bias in the animal’s choice sequences. As a result, theoretically, the maximum rate of reward that can be earned by the animal was 50%. Across all animals, the average reward rate ranged from 46.5% to 49.0%, indicating that they avoided making too many predictable choices.

The animal’s highly stochastic choices, however, could have still resulted from the use of a reinforcement learning algorithm. Consistent with this possibility, the animals indeed often displayed a small but significant bias to choose the same target rewarded in the previous trial, and to switch to the other target otherwise (Figure 2A; Lee et al., 2004; Seo and Lee, 2007). Referred to as a win-stay lose-switch (WSLS) strategy, this is naturally adopted by animals engaged in a wide variety of two alternative forced choice tasks (Sugrue et al., 2004; Dorris and Glimcher, 2004; Lau and Glimcher, 2005; Thevarajah et al., 2010). The WSLS strategy is suboptimal in the matching pennies task since it is exploited by the computer opponent. Despite this, all animals tested in this study adopted the WSLS strategy significantly more often than predicted by chance (Table 1). We also found that the tendency to repeat the choice rewarded in the previous trial was often stronger than the tendency to switch away from the unrewarded target (Table 1; Figure 2A). This suggests that the animal's behaviors might be better accounted for by a reinforcement learning algorithm in which the positive outcomes (wins) produce larger changes in the value function of the chosen action than negative outcomes (losses). To test this quantitatively, we applied a modified reinforcement learning model with separate reward parameters for win and loss outcomes (see Experimental Procedures). As expected, the results showed that the value function was updated to a greater degree in rewarded trials than in unrewarded trials (Figure 2B). We compared the goodness of fit (log likelihood) for this reinforcement learning model to that for the Nash equilibrium in which the probability of choosing each target was 0.5 in all trials. We found that the reinforcement learning model accounted for the animal's choices better than the equilibrium model in the majority of sessions (Table 1). For one of the animals (monkey D), the reinforcement learning model was better only in a minority of sessions (36.7%), but the same animal still displayed a stronger WS strategy than a LS strategy. Therefore, the behaviors of all animals were relatively close to, but still significantly deviated from the optimal strategy.

The dominance of win-stay strategy over lose-switch strategy was relatively stable both within a session as well as across sessions in each animal. When we compared the difference between the probabilities of win-stay vs. lose-switch strategies, or the win/loss reward parameters of the reinforcement learning model, for the first two blocks of 128 trials of the matching pennies task, we found no significant behavioral differences in any animals. We also found little evidence of systematic changes across recording sessions. The difference between the win-stay and lose-switch probabilities decreased significantly across daily sessions in only one animal (monkey C; correlation coefficient between p(win-stay) – p(lose-stay) and session number, r=−0.47, p=0.031). However, the difference in the win and loss reward parameters of the reinforcement learning models did not show any significant trend across recording sessions in any animals (p>0.15).

Reward Enhances Neural Signals Related to Previous Choices

Previous studies have shown that neurons in the DLPFC, LIP, and ACC often encoded the animal's choice and reward in the previous trial (Barraclough et al., 2004; Seo et al., 2007, 2009; Seo and Lee, 2007). In the present study, to gain further insights into how these choice and reward signals contributed to reinforcement learning, we tested whether and how cortical signals related to the animal's previous choices were influenced by previous reward during the matching pennies task. The following analyses were applied to the neurons recorded from the DLPFC (n=322), ACC (n=154), and LIP (n=205), which were described previously (Seo et al., 2007, 2009; Seo and Lee, 2007; Bernacchia et al., 2011), as well as the new dataset consisting of 185 neurons recorded from the SEF of two animals that had not been previously reported (see also Table S1).

Similar to the results from other cortical areas, we found that SEF neurons signaled the animal's choice and reward in the previous trial (Figure 3; see also Figure S1 for the results shown for different epochs). For example, during the 500-ms fore-period of the matching pennies task, a significant fraction of neurons in each area encoded the reward in the previous trial (Figure 3A). Among the neurons that showed significant effects of previous rewards, the proportion of neurons that increased their activity significantly when the animal was rewarded in the previous trial (SEF, 51.5%; DLPFC, 54.8%; LIP, 61.3%; ACC, 42.2%) was not significantly different from that of neurons decreasing their activity significantly following the rewarded trials, areas also changed their activity according to the choice of the animal in the previous trial (Figure 3B). In addition, a significant fraction of neurons in DLPFC, SEF, and LIP encoded an interaction between the animals' choice and reward in the previous trial, while ACC neurons lacked such interactions (Figure 3C; see also Figure S1). Overall, the proportion of neurons that significantly changed their activity during the cue period (500ms window from cue onset to fixation offset) according to the animal's choice in the previous trial, reward, and the interaction between the two was similar to the level observed for the fore-period (Figure 3). All the results described below were qualitatively similar for the fore-period and cue period when they were analyzed separately, so we report only the results from the analyses applied to the activity during the 1-s interval including both of these epochs.

The fact that many neurons in SEF, DLPFC, and LIP showed a significant interaction for the previous choice and reward implies that the signals related to the animal's choice in the previous trial was modulated or gated by the reward. In addition, during the matching pennies task, neurons in SEF, DLPFC, and LIP tended to encode the previous choice more robustly following rewarded trials than in unrewarded trials (Figure 4A–C). For example, for the SEF neuron shown in Figure 4A, the average firing rate during the cue period was 19.2 and 25.5 spikes/s, when the animal's leftward and rightward choices in the previous trial were rewarded, respectively. In contrast, the average firing rate was 18.8 and 21.3 spikes/s, following unrewarded leftward and rightward choices in the previous trial. The interaction between the previous choice and reward was highly significant for this neuron (t-test for the interaction term in a regression model, $p<10^{-8}$). During the matching pennies task, the animal's choices in two successive trials were largely uncorrelated since the tendency to repeat the same choices as well as the tendency to switch in two successive trials were exploited and penalized by the computer opponent. This allowed us to determine unequivocally whether receiving a reward in the previous trial enhanced encoding of the choice in the previous trial or the upcoming choice in the current trial. For example, for the neurons shown in Figure 4, the change in their activity related to the animal's upcoming choice was not significantly affected by the reward in the previous trial ($p>0.5$; Figure 4D–F).

To test whether signals related to the animal's previous choice were affected by reward consistently across the population, we applied a linear discriminant analysis separately to each neuron, and asked how well we could decode the animals' previous choice using the firing rates of the neurons in each trial separately according to whether the previous choice was rewarded or not. We found that the previous choice of the animals was more robustly encoded after receiving a reward in SEF (paired t-test, $p<0.01$), DLPFC ($p<10^{-4}$), LIP ($p<10^{-3}$), but not in ACC ($p=0.817$; Figure 5A). When we applied the same analysis for the animal's upcoming choice, the decoding accuracy was not consistently affected by whether the animal was rewarded in the previous trial or not. This difference in the decoding accuracy for the upcoming choice was not significant in any area: SEF (paired t-test, $p=0.516$), DLPFC ($p=0.632$), LIP ($p=0.183$), and ACC ($p=0.696$; Figure 5B). Thus, during the matching pennies task, information about previous but not upcoming choice was encoded more robustly in the SEF, DLPFC, and LIP after the animals received a reward.

Behavioral Correlates of SEF Activity Related to Previous Choices

The fact that the animals relied more on the win-stay strategy than on the lose-switch strategy during the matching pennies task indicates that there was an overall bias for the animal to choose the same target in successive trials (see also Figure S2A). Therefore, to test whether this behavioral bias could potentially result from the cortical activity related to the animal's previous choices, we estimated the probability for the animal to choose the same target in small blocks of trials (mean = 163 trials; see Experimental Procedures), and examined whether this was correlated with the difference in the decoding accuracies for the animal's previous choices after rewarded and unrewarded trials. We found that these two measures were significantly correlated only for the SEF. Namely, the degree of improved decoding accuracy for previously rewarded choice was significantly and positively correlated with the probability of switching for the SEF (r=0.14; $p<10^{-3}$), but not for the other cortical areas (DLPFC, r=0.01, LIP, r=0.05, ACC=−0.02; p>0.3). This positive correlation between switching probability and difference in decoding accuracy was statistically significant, even when the results were analyzed for the neurons from the SEF of each animal separately (monkey D: r=0.15, $p<10^{-3}$; monkey E: r=0.12; p<0.05), but not for any other cortical area (see also Figure S2B).

Although we found a significant correlation between the reward-related decoding accuracy of previous choices and switching behavior only for the SEF, the behaviors of the two animals in which the SEF activity was recorded (monkeys D and E) did not deviate substantially from those of other animals. For example, the probability of switching or the parameters of the reinforcement learning model estimated for these two animals were not extreme (see Figure S2A). We have also tested whether the increase or decrease in the probability of switching was correlated with the changes in the decoding accuracy of previous choices across two successive blocks of trials in the same session (see Experimental Procedures). Again, we found a significant correlation only in the SEF (r=0.16, $p<10^{-3}$), but not for the other cortical areas (DLPFC, r=0.03; LIP, r=0.00; ACC, r = 0.01; p>0.5). In the modified reinforcement learning model used to analyze the animal's behavior, the sum of the win and loss reward parameters ($\theta_{win}+\theta_{loss}$) quantifies the dominance of win-stay strategy over the lose-switch strategy, since this sum would be zero, if these two opposite strategies were applied equally. We found that the sum of the win and loss reward parameters was also significantly correlated with the difference in the decoding accuracy for previous choice in the SEF (r = −0.12, $p<10^{-3}$), but not in other cortical areas (DLPFC, r=−0.04; LIP, r=−0.02; ACC, r=0.05; p>0.2). Therefore, the signals related to the previous choices in the SEF might contribute to attenuating the tendency to choose the same target repeatedly in the matching pennies task.

Choice Signals during Visual Search Task

All the neurons analyzed in the present study were also tested during the visual search task in which correct target locations were explicitly signaled by visual cues (Figure 1B). We therefore examined neural activity related to previous choices and rewards during the search task, in order to test whether neural activity related to the animal's previous choices was affected by the task demands. Importantly, the timing of task events and the reward statistics were equated for both tasks. During the visual search task, one of the targets was red and the other was green, and the animals were required to choose the green target but rewarded in only 50% of the correct trials so that the reward rate was approximately equal for the two tasks. The location of the green target was controlled such that the animal was required to choose the target in the same location as in the previous trial in 50% of the trials regardless of whether the animal was rewarded or not. Thus, the choice and reward in a given trial did not have any predictive value for the choice in the next trial during the visual search task.

For the example neuron in the SEF shown in Figure 6, the activity for the monkey's previous choice was more robustly encoded during the matching pennies task, but there was little effect of reward on previous choice signals during the visual search task (3-way task × previous choice × previous reward interaction during cue-period: $p<10^{-8}$). To test whether the same tendency was present for the population of neurons in each area, we applied the linear discriminant analysis described above to all neurons separately for the two tasks.

Since there were always fewer trials for the visual search task than for matching pennies task, we randomly selected a subset of trials from the matching pennies task so that we could fairly compare the results between the two tasks. We found that during the visual search task, the reward-dependent enhancement of previous choice signals was relatively weak in all cortical areas, and statistically significant only in LIP (paired t-test, p=.033), but not in SEF (p=0.741), DLPFC (p=0.058), or ACC (p=0.258; Figure 7A). The statistically significant reward-dependent enhancement observed in the LIP might reflect a type I error, since it is no longer significant when corrected for multiple comparisons. We also found that the average decoding accuracy for the upcoming choice was not significantly affected by the reward in the previous trial in any cortical area (paired t-test, p>0.4).

To test more systematically whether the SEF was unique in encoding the animal's previous choice more reliably when it was rewarded during the matching pennies task, we applied a 3-way ANOVA with repeated measures on the decoding accuracies with cortical area as a between-subject variable and previous reward and task as within-subject variables. This analysis found a significant 3-way interaction between cortical area, task, and previous reward (p<0.05; Figure 7B). This 3-way interaction was driven by the results from the SEF. First, we performed a 2-way ANOVA for task and previous reward separately for each area and found that the SEF was the only area showing a significant interaction between task and previous reward (SEF, p<0.01; DLPFC, p=0.598; LIP, p=0.577; ACC, p=0.393). Second, we repeated the 3-way ANOVA for each pair-wise combination of areas. A significant 3-way interaction was found for all pairs including the SEF (p<0.05), but not for any pairs without the SEF. Thus, the SEF might play a unique role in encoding past events differently depending on the context of the particular task performed by the animal.

DISCUSSION

Cortical Encoding of Rewarded Actions

Persistent neural activity related to the animal's previous choices and rewards might play a role in reinforcement learning by linking past actions and rewards with future behavioral plans (Curtis and Lee, 2010; Lee et al., 2012a). The recent history of an animal's rewards modulates task-relevant activity in multiple regions of the primate and rodent brains, including DLPFC (Barraclough et al., 2004; Histed et al., 2009), ACC (Seo and Lee, 2007), LIP (Seo et al., 2009), striatum (Histed et al., 2009; Kim et al., 2009; Kim et al., 2013), orbitofrontal cortex (Sul et al., 2010; Kennerley et al., 2011), and hippocampus (Wirth et al., 2009; Singer and Frank, 2009; Lee et al., 2012b). Such persistent signals might contribute to more robust encoding of certain task-specific events in rewarded trials. Similarly, sharp wave ripple activity related to the reactivation of past experiences in hippocampal place cells is more robust if the animal was previously rewarded (Singer and Frank, 2009), suggesting that they might also contribute to linking the memories of a given action and its consequences.

Previous studies have suggested that rewards might enhance the neural signals related to upcoming choices. For example, the single-neuron activity related to the animal's upcoming choices was enhanced in the DLPFC and caudate nucleus during a paired-association task when the animal was rewarded in the previous trial (Histed et al., 2009). Our results suggest that rewards lead to a more robust encoding of past actions, but have little effect on the encoding of upcoming actions in multiple cortical areas, including the DLPFC. The discrepancy between our findings and previous work might reflect a difference in the behavioral tasks. For example, during the matching pennies task, animals displayed weak but significant biases to use the win-stay-lose-switch strategy, although this was exploited by the computer opponent and led to suboptimal outcomes. In the associative learning task (Histed et al., 2009), animals learned a correct action associated with each visual stimulus over a large number of trials, and the identities of visual stimuli and hence correct actions were randomized across trials. This might affect how the brain retrieves the information about the previously rewarded choices. In addition, behavioral tasks used in neurophysiological studies often require animals to associate either spatial locations or other visual features with rewards. Since such associations are learned over many trials, it is often difficult to disambiguate neural signals related to past and future events (Baeg et al., 2003). The matching pennies task is unique in this respect in that successive choices freely made by the animal are largely independent. This allowed us to show clearly that rewards enhanced the fidelity of neural signals related to past, rather than upcoming actions.

The animal's behavior during the matching pennies was driven more by a win-stay strategy than a lose-switch strategy. The analysis of the animal's choice behavior based on a reinforcement learning model also revealed that the animals made greater adjustments to their future behavior after their choices were rewarded compared to when they were not rewarded. Such asymmetric effects of rewarding and non-rewarding (or punishing) outcomes have been described previously under various contexts (Nakatani et al., 2009; Kravitz et al., 2012). This might be expected from an ecological perspective. In the real-world environment, animals face a large number of possible actions and only infrequently receive reward. Therefore, rewarding outcomes might be more informative than non-rewarding outcomes. Currently, how the information about reward and penalty is processed in the brain and influences the animal's future behaviors remains poorly understood. It has been proposed that the direct and indirect pathways in the basal ganglia might be specialized in reinforcement and punishment (Frank et al., 2004; Kravitz et al., 2012). On the other hand, rewarding and punishing outcomes are often processed by the same neurons in the primate frontal cortex (Seo et al., 2009). Similarly, although dopamine neurons encode signals related to unexpected rewards as well as unexpected penalties (Matsumoto and Hikosaka, 2009), how such signals contribute to updating the values of different actions needs to be investigated further.

Role of SEF in Exploration and Strategic Adjustment

We found that signals related to previously rewarded actions were enhanced compared to those related to unrewarded actions in the SEF, DLPFC, and LIP, but this difference was significantly correlated with the animal's behavior only for the SEF. Specifically, the animals were more likely to choose the target not chosen in the previous trial during a block of trials in which the SEF activity encoded the previously rewarded actions more reliably. In the context of the matching pennies task, this might have improved the animal's performance by facilitating exploratory behavior. The matching pennies task is designed to discourage the animals from adopting stereotypical sequences of choices and to encourage them to make their choices randomly. Nevertheless, the animals tended to adopt the win-stay-lose-switch strategy, even though this reduced their overall reward rate. The SEF may provide top-down control to counteract the biases resulting from such simple reinforcement learning and contribute to exploratory switching behavior. The possibility that the signals in the SEF related to the animal's previous choice are actively maintained and influences the animal's decision making is also supported by the fact that such signals were significantly attenuated during the visual search only in the SEF, but not in other cortical areas. In the visual search task, the animals learned a simple rule by which only a particular cue was associated with reward, and this rule was fixed throughout the entire experiment. In contrast, the matching pennies task put the animals in a perpetual state of exploration. The stimuli were identical on every trial and the animals had to continually explore and avoid developing strong stimulus-response associations in order to maximize rewards. Although we tested all the neurons in the visual search task before they were tested in the matching pennies task, the fact that reward-dependent signals related to previous choice were attenuated during the visual search only in the SEF makes it unlikely that this was related to time-dependent changes in motivational factors. Therefore, SEF may play a unique role in exploration or suppressing the influence of fixed stimulus-response associations.

Consistent with the possible role of SEF in exploration, a previous study showed that a human subject with a focal lesion to SEF displayed specific deficits in switching behavior (Parton et al., 2007). In that study, the subject was cued by the color of a fixation cross to make a saccade either towards a target (pro-saccade) or in the opposite direction of the target (antisaccade). Thus, the subject had to learn to choose between two different stimulus-response mappings. While the subject had no difficulty performing either pro or anti-saccades in themselves, he made more errors when required to switch from anti-saccades to pro-saccades. In addition, the same subject was impaired in rapidly updating his saccade plans, without showing any impairment in detecting errors after making saccades to incorrect targets (Husain et al., 2003). Based on these results, it was suggested that the SEF might be important in providing top-down control signals when rules governing stimulus-response mappings are in conflict. Therefore, these results suggest that the SEF might be involved in exploratory behaviors regardless of the exact nature of switching, namely, whether the switch occurs among different spatial target locations or different behavioral rules.

Previous studies have suggested that the medial frontal cortex, including the SEF, plays a more important role for behaviors that are guided by internal cues, rather than directly by incoming sensory stimuli (Tanji, 1996). For example, neural activity related to upcoming movements often occur earlier in medial frontal regions, such as the SEF and supplementary motor area, than in other areas (Coe et al., 2002; Haggard, 2008; Sul et al., 2011). In addition, it was found in a recent study that very few neurons in SEF signaled target locations during a visual search task or exhibited priming effects during pop-out (Purcell et al., 2012). Instead, neurons in SEF strongly modulated their activity following errant saccades. Consistent with the results obtained from other tasks, such as anti-saccade (Schlag-Rey et al., 1997) and saccade countermanding task (Stuphorn et al., 2000), these results indicate that the SEF might be involved when a greater degree of cognitive control is necessary for optimal performance. Other studies have also shown that the SEF is only loosely related to direct transformations between sensory inputs and motor outputs. For example, neurons in the SEF can also change their directional tuning while the animals acquire novel stimulus-response mappings (Chen and Wise, 1995, 1996, 1997). SEF has also been shown to encode information differently depending on visual context (Olson and Gettner, 1995; Olson et al., 2000) or in anti-saccade tasks (Schlag-Rey et al., 1997). Additionally, SEF robustly signals errors or rewards after saccades (Stuphorn et al., 2000; Stuphorn et al., 2010) and has also been shown to contain signals relevant for metacognition (Middlebrooks and Sommer, 2012), suggesting that this area may play a role in monitoring the performance of the animal, possibly to bias future behavior appropriately according to past events. Although we focused on the SEF in the present study, similar signals might also exist in the nearby regions in the medial frontal cortex, including the pre-supplementary motor area (Isoda and Hikosaka, 2007).

Cortical Mechanisms of Reinforcement Learning

Signals related to the animal's previous choices in the SEF might uniquely contribute to the animal's decision making strategies, since they were correlated with the animal's subsequent choices during the matching pennies task and also attenuated during the visual search. Nevertheless, the neural activity in all cortical areas tested in the present study displayed several common properties. For example, signals related to the animal's previous choices persisted for several trials not only in the SEF, but also in the DLPFC and LIP (Seo et al., 2007, 2009). In contrast, choice signals were relatively weak and decayed more quickly in the ACC, whereas the signals related to reward were more robust and persistent in the ACC (Seo and Lee, 2007, 2008). However, it should be emphasized that the signals related to the previous rewards were present in all of these cortical areas. Moreover, signals related to rewards were found in the rodent cortical and subcortical areas (Sul et al., 2010, 2011; Lee et al., 2012b; Kim et al., 2013) as well as practically throughout the entire human brain (Vickery et al., 2011). In addition, neurons encoding specific conjunctions of actions and rewards are found in multiple cortical areas, although they were more robust in the SEF than in the DLPFC and ACC (Seo and Lee, 2009). Such conjunctive signals are thought to contribute to the process of updating the values of specific actions according to the animal's experience (Lee et al., 2012a). These results suggest that the signals necessary for encoding rewards and upgrading the value function of chosen action are available in a large number of brain areas.

The widespread presence of signals related to previous choices and rewards does not imply that they serve the same functions in different brain areas. To the contrary, signals in DLPFC, LIP, and ACC have been shown to play a diverse role in dynamically encoding and updating values associated with different parameters related to the environment as well as the animal's own reward and action history (Seo and Lee, 2008, 2009; Rushworth and Behrens, 2008; Wallis and Kennerley, 2010). For example, recent studies in ACC have shown that while choice-related information is not encoded very strongly, it plays a key role in representing and learning positive and negative values of actual, potential, and hypothetical outcomes (Brown and Braver, 2005; Seo and Lee, 2007; Quilodran et al., 2008; Kennerley and Wallis, 2009; Hayden et al., 2009), as well as adjusting learning rates as a function of environmental volatility (Behrens et al., 2007). LIP has also been implicated in value representation (Platt and Glimcher, 1999; Sugrue et al., 2004; Seo et al., 2009) as well as sensory evidence accumulation in perceptual decision-making tasks (Roitman and Shadlen, 2002). The fact that multiple cortical areas tend to encode past actions more robustly following rewarded trials suggests that they may be integrating information related to the animal's past events to guide future actions, perhaps biasing the animals to rely on a simple, model-free reinforcement learning algorithm. When this leads to a suboptimal outcome, as during a competitive social interaction, the medial frontal areas, such as the SEF, might play an important role in detecting and overriding such a default reinforcement learning strategy. It is possible that some of other brain areas might also make unique and specific contributions according to the demands of specific tasks. Future studies should explore this by using dynamic task designs that encourage animals to combine information from past events and newly available sensory information and use this information to guide their future choices.

EXPERIMENTAL PROCEDURES

Animal Preparation

Six rhesus monkeys (5 male, C, D, E, H, and I; 1 female, K; body weight = 5~12 kg) were used. Eye movements were monitored at a sampling rate of 225 Hz with a high-speed video-based eye tracker (ET49; Thomas Recording). Some of the procedures were performed at the University of Rochester and were approved by the University of Rochester Committee on Animal Research. The rest of the procedures were approved by the Institutional Animal Care and Use Committee (IACUC) at Yale University.

Behavioral Tasks

The data analyzed in this study were obtained from two different oculomotor tasks: a visual search task and a free choice matching pennies task. During the visual search task, trials began when the animals fixated a small yellow square at the center of a computer monitor for 0.5s (fore-period). Next, a pair of green and red disks (radius=0.6°) were presented along the horizontal meridian (eccentricity = 5°). The fixation target was extinguished after a 0.5s delay (cue-period), and the animal was required to shift its gaze towards the green target. After maintaining fixation on the chosen target for 0.5, a red ring appeared around the green target (feedback period) and the animal was rewarded randomly with a 50% probability in correct trials either 0.2s (DLPFC recording in monkeys C and E, Barraclough et al., 2004) or 0.5s (all other monkeys/regions) later. The location of the green target was selected pseudo-randomly such that it was equally likely to appear at the left and right locations. In addition, the reward and location of the green target in 3 consecutive trials were fully balanced such that each of the 64 possible combinations of target locations and rewards in a 3 trial sequence was presented twice. Since choices and rewards of previous trials cannot be defined in the beginning of each session, this was padded by 2 additional trials, resulting in a total of 130 search trials.

During the matching pennies task, the animals played a competitive game against a computer opponent (Barraclough et al., 2004; Lee et al., 2004; Seo and Lee, 2007; Seo et al., 2009). The animal was presented with two identical green targets during the cue period, but otherwise, the timing of events during the matching pennies task was identical to that of the visual search task. The animal was rewarded only if it chose the same target as the computer, which was indicated by the red ring during the feedback period. The computer was programmed to exploit statistical biases in the animal's behavior by analyzing the animal's choice and reward history (algorithm 2 in Lee et al., 2004). For example, if the animal chose the left target more frequently, the computer opponent increased the frequency of choosing the right target. Similarly, if the animal tended to choose the same target rewarded in the previous trial (i.e., win-stay strategy), this tendency was also exploited by the computer opponent. As a result, the matching pennies task encouraged the animal to make its choices randomly and independently across trials, which is advantageous for dissociating neural signals related to previous and current trials and other reward-dependent strategies (Lee and Seo, 2007). In each recording session, neural activity was first recorded while the monkeys performed the visual search task. Neurons were included in the analyses only if they were tested in at least 130 trials for both tasks.

Analysis of Behavioral Data

We analyzed a number of probabilities related to the animal's strategies and rewards using a series of binomial tests. For example, whether the animal chose the two targets equally often was tested with a two-tailed binomial test, separately for each session as well as for all the sessions from each animal. We also used a one-tailed binomial test to determine whether the number of sessions with significant bias for one of the targets was significantly higher than expected by chance. Similarly, whether the win-stay-lose-switch strategy was used more frequently than by chance (p=0.5) was determined with a one-tailed binomial test. Whether the proportion of trials in which the animal applied a win-stay strategy after rewarded trials, p(winstay), was significantly different from the proportion of trials in which the animal applied a lose-switch strategy after unrewarded trials, p(lose-switch), was tested using a two-proportion z-test separately for each session as well as for all sessions combined for each animal. For all hypothesis testing, the significance level of 0.05 was used.

The choice data from each animal were also analyzed using a modified reinforcement learning model. In this model, the value functions for the left and right targets in a trial, Q_t(left) and Q_t(right), were updated differently depending on whether the animal's choice was rewarded or not. Namely, the value function for target x on trial t, $Q_t(x)$, was updated according the following equation,

$$\mathrm{Q_{t+1}(x)=\gamma Q_t(x)+W_t(x),}$$

where γ is a decay parameter. $W_t(x)$ indicates the additional changes in the value function when the target x was chosen in trial t. Namely, $W_t(x) = \theta_{win}$ if the target x was chosen and rewarded, θ_{loss} if the target x was chosen and unrewarded, and 0 otherwise. The probability of choosing each target was then given by the logistic transformation of the difference in the value functions for the two choices.

$$\mathrm{logit p_t(right) \equiv log p_t(right)/p_t(left)=Q_t(left) - Q_t(right).}$$

If the rewarded and unrewarded trials have equal and opposite influence on the animal's subsequent choices, the sum of θ_{win} and θ_{loss} should be equal to 0. Therefore, $\theta_{win} + \theta_{loss}$ reflects whether the animal's choice is influenced more strongly by the rewarding or non-rewarding outcomes in previous trials. Model parameters were estimated separately for each recording session using the maximum likelihood method (Pawitan, 2001).

Neurophysiological Recording

Single-unit activity was recorded from neurons in four different cortical regions (ACC, DLPFC, LIP, and SEF) of six monkeys using a five-channel multi-electrode recording system (Thomas Recording, Giessen, Germany). The methods used to localize the neurons in DLPFC (Barraclough et al., 2004), LIP (Seo et al., 2009), and ACC (Seo and Lee, 2007) have been previously described. Briefly, all neurons in the DLPFC (monkeys C, E, H, I, and K) were located anterior to the frontal eye field, which defined by eye movements evoked by electrical stimulation (current<50μA). Similarly, neurons were localized in the SEF (monkeys D and E) using electrical stimulation (current<100μA; see also Figure S3; Schlag and Schlag-Rey, 1987). All neurons in the ACC were recorded from the dorsal bank of the cingulate sulcus, directly beneath the recording sites in the SEF. In LIP (monkeys H, I, and K), neurons were recorded at least 2.5mm below the cortical surface along the lateral bank of the intraparietal sulcus. All of the datasets except for the data from SEF have been previously described (ACC - Seo and Lee, 2007; DLPFC - Barraclough et al., 2004; Seo et al., 2007; LIP - Seo et al., 2009).

Analysis of Neural Data

Trial-by-trial Analysis: Previous Choice Signals—Whether the activity during the fore-period and cue period significantly encoded the animal's previous choice, previous reward, and their interaction (Figure 3) was determined with the following regression model,

$$\mathrm{y(t)=a_0+a_1Ch(t)+a_2Ch(t-1)+a_3R(t-1)+a_4Ch(t-1) \times R(t-1),}$$

where Ch(t) and R(t) indicate the animal's choice (−1 and 1 for left and right choices, respectively) and reward (−1 and 1 for unrewarded and rewarded trials, respectively) in trial t, and a_0~a_4 the regression coefficient. The statistical significance of each regressor was determined with a t-test.

To investigate how reliably information about the animal's choice in the previous or current trial could be decoded from the neural activity in a given trial, we applied a linear discriminant analysis with 5-fold cross validation to each neuron separately. To build an unbiased classifier, we balanced the dataset within each neuron over all possible combinations of the animal's previous and current choice, and the reward in the previous trial, by randomly removing trials until each category contained the same number of trials. First, we decoded the previous choice of the animal separately following rewarded and unrewarded trials (Figure 5A). To examine the time course of this information, the linear discriminant analysis was applied with a 500-ms sliding time window advancing in 50ms increments. For cross validation, trials within each of 8 categories were randomly assigned to 5 different subgroups (5-fold cross validation). For each subgroup, we used the trials in the 4 remaining subgroups as a training set to find the firing rate boundary that best classified the animal's previous choice. We then determined how well we could classify the subgroup's trials using this boundary. Cross-validation was performed by repeating this procedure for each sub-group and averaging the results. Population means were constructed by averaging these values for all neurons within each region. Decoding of current choice activity (Figure 5B) was obtained similarly. For comparison across task types (matching pennies versus visual search; Figure 7), we randomly removed trials from the matching pennies task so that each task contained the same number of trials before performing the discriminant analysis. Statistical tests between previously rewarded and non-rewarded trials were performed with a paired t-test, after first transforming the decoding accuracy using the arcsine function. However, results were unaffected when analyzed without the arcsine transformation.

To test whether the difference in the decoding accuracy after rewarded vs. unrewarded outcomes varied significantly between the visual search and matching pennies and across different cortical areas, a 3-way analysis of variance (ANOVA) with repeated measures was used (reward × task ×cortical area) with reward and task as within-subject variables. This analysis was applied to the decoding accuracy estimated with a linear classifier applied to the activity during a 1-s window comprising the fore and cue-periods (−0.5s to 0.5s relative to target onset). In order to determine which areas contributed to the observed significant 3-way interaction, we compared the influence of different regions by running the same 3-way ANOVA between all pair-wise subsets of cortical regions. We also ran a 2-way ANOVA (reward × task) separately on each cortical region.

Block-wise Analysis: Neural-Behavioral Correlation—When the proportions of rewarded and unrewarded trials are similar, the difference in the frequencies of using the win-stay vs. lose-switch strategy, namely, p(WS)–p(LS) is approximately equal to the probability of choosing the same target in two successive choices, p(stay).Therefore, to test whether the difference in the decoding accuracy following rewarded vs. unrewarded outcomes was related to the animal's behavioral strategy, we examined the correlation between p(switch) = 1–p(stay) and the difference in the decoding accuracy after rewarded and unrewarded outcomes. To increase statistical power, we divided all recording sessions into several blocks of trials. Each block contained 12 trials for each of the 8 different trial types defined by the animal's previous and current choices, and the reward for the previous choice. These blocks were created by scanning through each session until each of the 8 categories was populated by at least 12 trials (a total of 96 trials in each block), and the same decoding analysis described above was performed. The probability of stay was calculated on the entire block (mean±SD = 163.3±38.2 trials). Therefore, the neural decoding analysis was performed on a subset of the same trials used to calculate the behavioral measure. Any blocks near the end of sessions which did not contain at least 12 trials per category were removed from the analysis. The modified reinforcement learning model was also run separately on the exact same blocks of trials that were used for calculating p(switch). We also examined whether changes in p(switch) between two successive blocks was correlated with the difference in decoding accuracy. For these analyses, we only examined data in which neurons were held for 2 or more blocks (165, 187, 100, and 135 neurons in the SEF, DLPFC, LIP, and ACC, respectively).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dominic Barraclough and Ben McGreevy for their help with data collection. This study was supported by the National Institute of Drug Abuse (R01 DA029330), and CD was supported by a training grant from the National Institute of Neurological Disorders and Stroke (T32 NS007224).

REFERENCES

Aston-Jones G, Cohren JD. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. Annu. Rev. Neurosci. 2005; 28:403–450. [PubMed: 16022602]

Averbeck BB, Sohn JW, Lee D. Activity in prefrontal cortex during dynamic selection of action sequences. Nat. Neurosci. 2006; 9:276–282. [PubMed: 16429134]

Baeg EH, Kim YB, Huh K, Mook-Jung I, Kim HT, Jung MW. Dynamics of population code for working memory in the prefrontal cortex. Neuron. 2003; 40:177–188. [PubMed: 14527442]

Barraclough DJ, Conroy ML, Lee D. Prefrontal cortex and decision making in a mixed-strategy game. Nat. Neurosci. 2004; 7:404–410. [PubMed: 15004564]

Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS. Learning the value of information in an uncertain world. Nat. Neurosci. 2007; 10:1214–1221. [PubMed: 17676057]

Bernacchia A, Seo H, Lee D, Wang XJ. A reservoir of time constants for memory traces in cortical neurons. Nat. Neurosci. 2011; 14:366–372. [PubMed: 21317906]

Brown JW, Braver TS. Learned prediction error likelihood in the anterior cingulate cortex. Science. 2005; 307:1118–1121. [PubMed: 15718473]

Chen LL, Wise SP. Neuronal activity in the supplementary eye field during acquisition of conditional oculomotor associations. J. Neurophysiol. 1995; 73:1101–1121. [PubMed: 7608758]

Chen LL, Wise SP. Evolution of directional preferences in the supplementary eye field during acquisition of conditional oculomotor associations. J. Neurosci. 1996; 16:3067–3081. [PubMed: 8622136]

Chen LL, Wise SP. Conditional oculomotor learning: population vectors in the supplementary eye field. J. Neurophysiol. 1997; 78:1166–1169. [PubMed: 9307145]

Coe B, Tomihara K, Matsuzawa M, Hikosaka O. Visual and anticipatory bias in three cortical eye fields of the monkey during an adaptive decision-making task. J. Neurosci. 2002; 22:5081–5090. [PubMed: 12077203]

Curtis CE, Lee D. Beyond working memory: the role of persistent activity in decision making. Trends Cogn. Sci. 2010; 14:216–222. [PubMed: 20381406]

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans. Nature. 2006; 441:876–879. [PubMed: 16778890]

Dorris MC, Glimcher PW. Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. Neuron. 2004; 44:365–378. [PubMed: 15473973]

Frank MJ, Seeberger LC, O'Reilly RC. By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science. 2004; 306:1940–1943. [PubMed: 15528409]

Haggard P. Human volition: towards a neuroscience of will. Nat. Rev. Neurosci. 2008; 9:934–946. [PubMed: 19020512]

Hayden BY, Pearson JM, Platt ML. Fictive reward signals in the anterior cingulate cortex. Science. 2009; 324:948–950. [PubMed: 19443783]

Histed MH, Pasupathy A, Miller EK. Learning substrates in the primate prefrontal cortex and striatum: sustained activity related to successful actions. Neuron. 2009; 63:244–253. [PubMed: 19640482]

Husain M, Parton A, Hodgson TL, Mort D, Rees G. Self-control during response conflict by human supplementary eye field. Nat. Neurosci. 2003; 6:117–118. [PubMed: 12536212]

Isoda M, Hikosaka O. Switching from automatic to controlled action by monkey medial frontal cortex. Nat. Neurosci. 2007; 10:240–248. [PubMed: 17237780]

Kennerley SW, Wallis JD. Neurons in the frontal lobe encode the value of multiple decision variables. J. Cogn. Neurosci. 2009; 27:8366–8377.

Kennerley SW, Behrens TEJ, Wallis JD. Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. Nat. Neurosci. 2011; 14:1581–1589. [PubMed: 22037498]

Kim H, Lee D, Jung MW. Signals for previous goal choice persist in the dorsomedial, but not dorsolateral striatum of rats. J. Neurosci. 2013; 33:35–51. [PubMed: 23283320]

Kim H, Sul JH, Huh N, Lee D, Jung MW. Role of striatum in updating values of chosen actions. J. Neurosci. 2009; 29:14701–14712. [PubMed: 19940165]

Kravitz AV, Tye LD, Kreitzer AC. Distinct roles for direct and indirect pathway striatal neurons in reinforcement. Nat. Neurosci. 2012; 15:816–818. [PubMed: 22544310]

Lau B, Glimcher PW. Dynamic response-by-response models of matching behavior in rhesus monkeys. J. Exp. Anim. Behav. 2005; 84:555–579.

Lee D, Conroy ML, McGreevy BP, Barraclough DJ. Reinforcement learning and decision making in monkeys during a competitive game. Cogn. Brain Res. 2004; 22:45–58.

Lee D, Seo H. Mechanisms of reinforcement learning and decision making in the primate dorsolateral prefrontal cortex. Ann. N. Y. Acad. Sci. 2007; 1104:108–122. [PubMed: 17347332]

Lee D, Seo H, Jung MW. Neural Basis of Reinforcement Learning and Decision Making. Annu. Rev. Neurosci. 2012a; 35:287–308. [PubMed: 22462543]

Lee H, Ghim J-W, Kim H, Lee D, Jung MW. Hippocampal neural correlates for values of experienced events. J. Neurosci. 2012b; 32:15053–15065. [PubMed: 23100426]

Matsumoto M, Hikosaka O. Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature. 2009; 459:837–841. [PubMed: 19448610]

Middlebrooks PG, Sommer MA. Neuronal correlates of metacognition in primate frontal cortex. Neuron. 2012; 75:517–530. [PubMed: 22884334]

Nakatani Y, Matsumoto Y, Mori Y, Hirashima D, Nishino H, Arikawa K, Mizunami M. Why the carrot is more effective than the stick: different dynamics of punishment memory and reward memory and its possible biological basis. Neurobiol. Learn. Mem. 2009; 92:370–380. [PubMed: 19435611]

Nash JF. Equilibrium points in n-person games. Proc. Natl. Acad. Sci. USA. 1950; 36:48–49. [PubMed: 16588946]

Olson CR, Gettner SN. Object-centered direction selectivity in the macaque supplementary eye field. Science. 1995; 269:985–988. [PubMed: 7638625]

Olson CR, Gettner SN, Ventura V, Carta R, Kass RE. Neuronal activity in macaque supplementary eye field during planning of saccades in response to pattern and spatial cues. J. Neurophysiol. 2000; 84:1369–1384. [PubMed: 10980010]

Parton A, Nachev P, Hodgson TL, Mort D, Thomas D, Ordidge R, Morgan PS, Jackson S, Rees G, Husain M. Role of the human supplementary eye field in the control of saccadic eye movements. Neuropsychologia. 2007; 45:997–1008. [PubMed: 17069864]

Pawitan, Y. In all likelihood: statistical modeling and inference using likelihood. Oxford: Oxford University Press; 2001.

Platt ML, Glimcher PW. Neural correlates of decision variables in parietal cortex. Nature. 1999; 400:233–238. [PubMed: 10421364]

Purcell BA, Weigand PK, Schall JD. Supplementary eye field during visual search: salience, cognitive control, and performance monitoring. J. Neurosci. 2012; 32:10273–10285. [PubMed: 22836261]

Quilodran R, Rothe M, Procyk E. Behavioral shifts and action valuation in the anterior cingulate cortex. Neuron. 2008; 57:314–325. [PubMed: 18215627]

Roitman JD, Shadlen MN. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. J. Neurosci. 2002; 22:9475–9489. [PubMed: 12417672]

Rushworth MFS, Behrens TEJ. Choice, uncertainty and value in prefrontal and cingulate cortex. Nat. Neurosci. 2008; 11:389–397. [PubMed: 18368045]

Schlag J, Schlag-Rey M. Evidence for a supplementary eye field. J. Neurophysiol. 1987; 57:179–200. [PubMed: 3559671]

Schlag-Rey M, Amador N, Sanchez H, Schlag J. Antisaccade performance predicted by neuronal activity in the supplementary eye field. Nature. 1997; 390:398–401. [PubMed: 9389478]

Seo H, Barraclough DJ, Lee D. Lateral intraparietal cortex and reinforcement learning during a mixed-strategy game. J. Neurosci. 2009; 29:7278–7289. [PubMed: 19494150]

Seo H, Lee D. Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. J. Neurosci. 2007; 27:8366–8377. [PubMed: 17670983]

Seo H, Lee D. Cortical mechanisms for reinforcement learning in competitive games. Philos. Trans. R. Soc. Lond. B. Bio. Sci. 2008; 363:3845–3857. [PubMed: 18829430]

Seo H, Lee D. Behavioral and neural changes after gains and losses of conditioned reinforcers. J. Neurosci. 2009; 29:3627–3641. [PubMed: 19295166]

Singer AC, Frank LM. Rewarded outcomes enhance reactivation of experience in the hippocampus. Neuron. 2009; 64:910–921. [PubMed: 20064396]

Stuphorn V, Taylor TL, Schall JD. Performance monitoring by the supplementary eye Field. Nature. 2000; 408:857–860. [PubMed: 11130724]

Stuphorn V, Brown JW, Schall JD. Role of supplementary eye field in saccade initiation: executive, not direct, control. J. Neurophysiol. 2010; 103:801–816. [PubMed: 19939963]

Sugrue LP, Corrado GS, Newsome WT. Matching behavior and the representation of value in the parietal cortex. Science. 2004; 304:1782–1787. [PubMed: 15205529]

Sul JH, Kim H, Huh N, Lee D, Jung MW. Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. Neuron. 2010; 66:449–460. [PubMed: 20471357]

Sul JH, Jo S, Lee D, Jung MW. Role of rodent secondary motor cortex in valuebased action selection. Nat. Neurosci. 2011; 14:1202–1208. [PubMed: 21841777]

Sutton, RS.; Barto, AG. Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press; 1998.

Tanji J. New concepts of the supplementary motor area. Curr. Opin. Neurobiol. 1996; 6:782–787. [PubMed: 9000016]

Thevarajah D, Webb R, Ferrall C, Dorris MC. Modeling the value of strategic actions in the superior colliculus. Front. Behav. Neurosci. 2010; 3:57. [PubMed: 20161807]

Vickery TJ, Chun MM, Lee D. Ubiquity and specificity of reinforcement signals throughout the human brain. Neuron. 2011; 72:166–177. [PubMed: 21982377]

Walllis JD, Kennerley SW. Heterogeneous reward signals in prefrontal cortex. Curr. Opin. Neurobiol. 2010; 20:191–198. [PubMed: 20303739]

Wirth S, Avsar E, Chiu CC, Sharma V, Smith AC, Brown E, Suzuki WA. Trial outcome and associative learning signals in the monkey hippocampus. Neuron. 2009; 61:930–940. [PubMed: 19324001]

Yu AJ, Dayan P. Uncertainty, Neuromodulation, and Attention. Neuron. 2005; 46:681–692. [PubMed: 15944135]

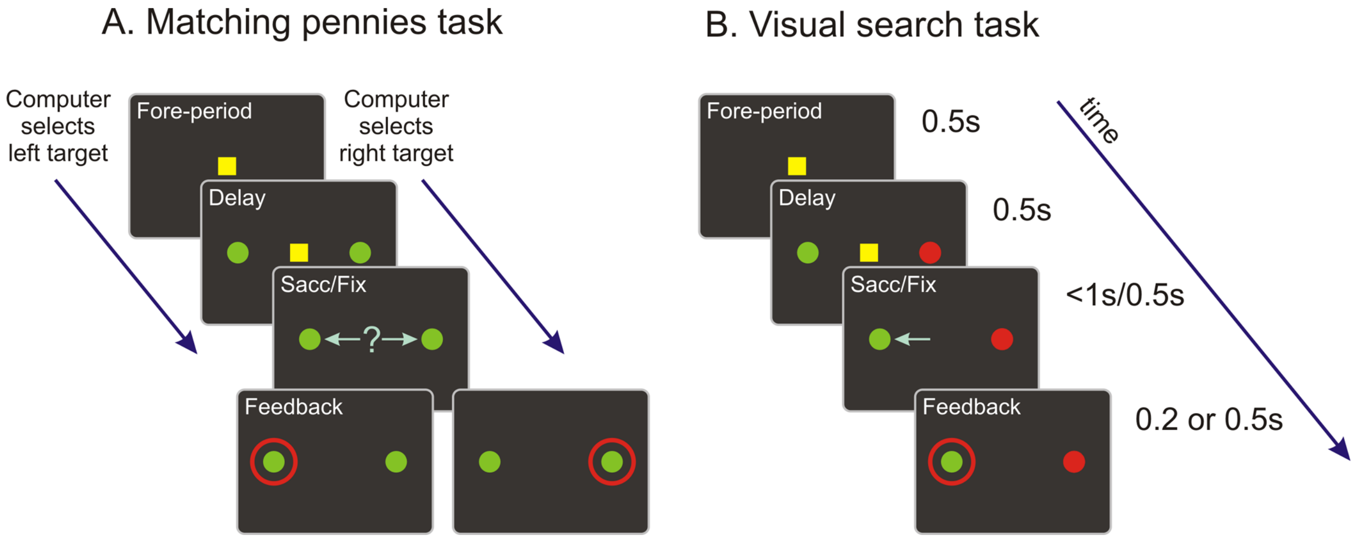

Figure 1.

Behavioral task. **A**. Matching pennies. **B**. Visual search tasks. The timing for each epoch was identical for the two tasks.

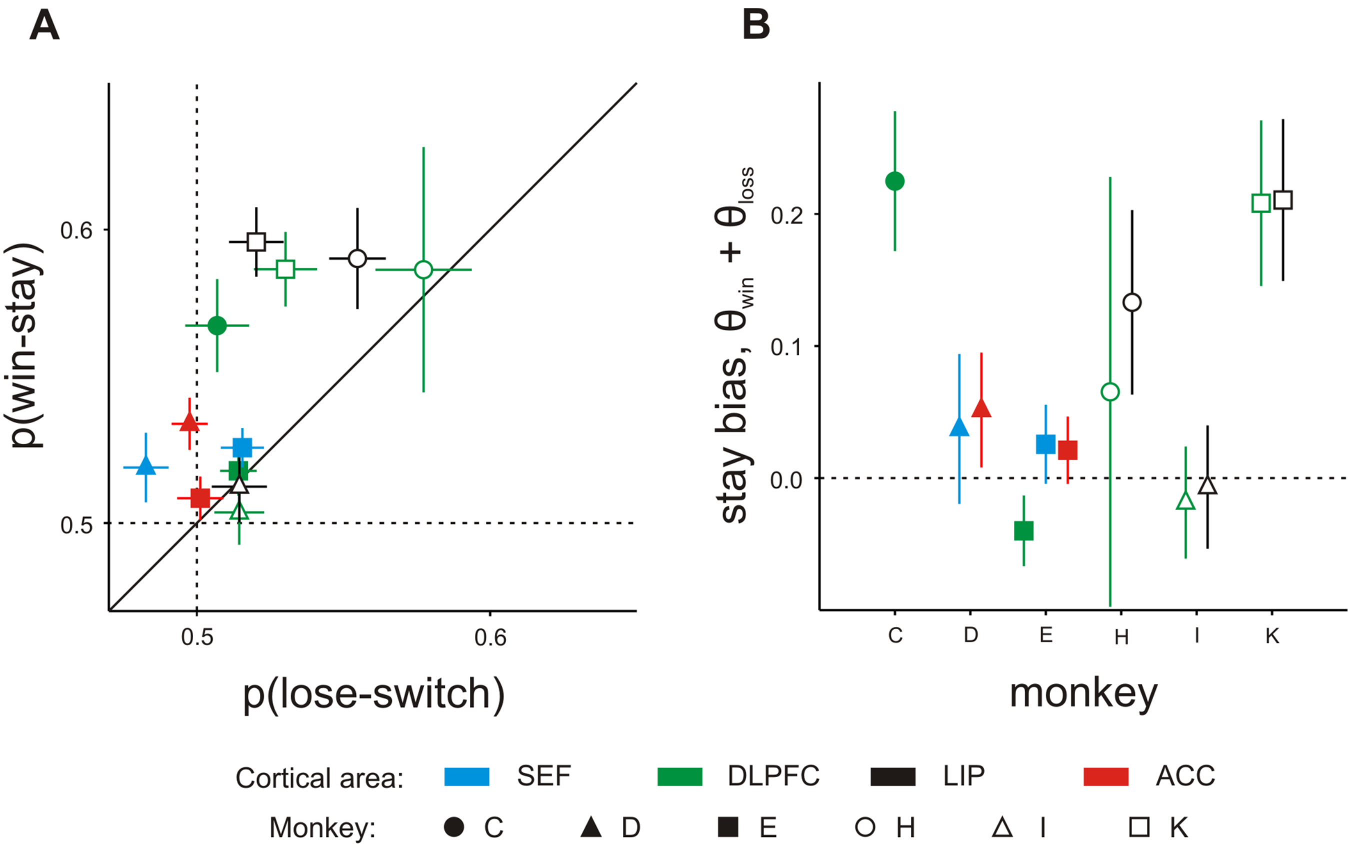

Figure 2.

Reinforcement learning during matching pennies task. **A**. Win-stay versus loseswitch behavior. The fraction of trials in which the monkeys chose the same target as in the previous trial after receiving a reward, p(win-stay), is plotted on the ordinate, and the fraction of trials in which they switched to the other target after not receiving a reward, p(lose-switch), is plotted on the abscissa. Different symbols and colors indicate different animals and cortical regions, respectively. **B**. Asymmetric effects of reward vs. no-reward estimated by a modified reinforcement learning model. Values for θ_{win} +θ_{loss} that are greater than zero indicate that rewarded trials had a greater effect on the animal's future behavior than non-rewarded trials. Error bars indicate mean ± SEM. See also Table S1.

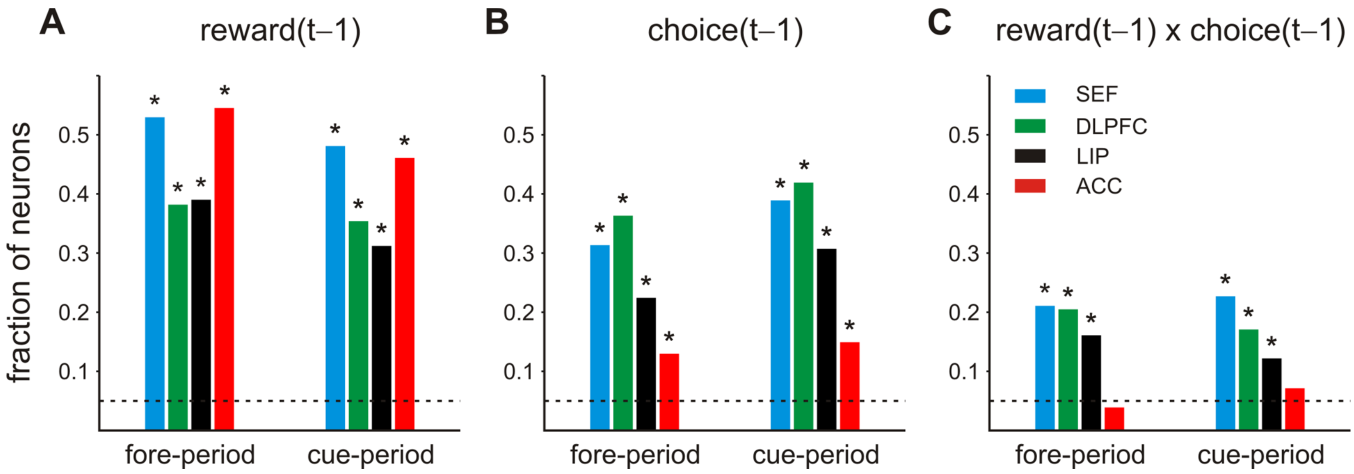

Figure 3.

Cortical signals related to previous choice and reward during the matching pennies task. The histograms show the fraction of neurons that modulated their activity significantly according to the previous outcome (rewarded or unrewarded, **A**), previous choice (left or right, **B**), or their interaction (**C**) in the fore-period and cue-period of the matching pennies task. Colors indicate different cortical regions. Dotted lines correspond to the significance level used (p=0.05). Asterisks, p<0.05 (binomial test). See also Figures S1 and S3.

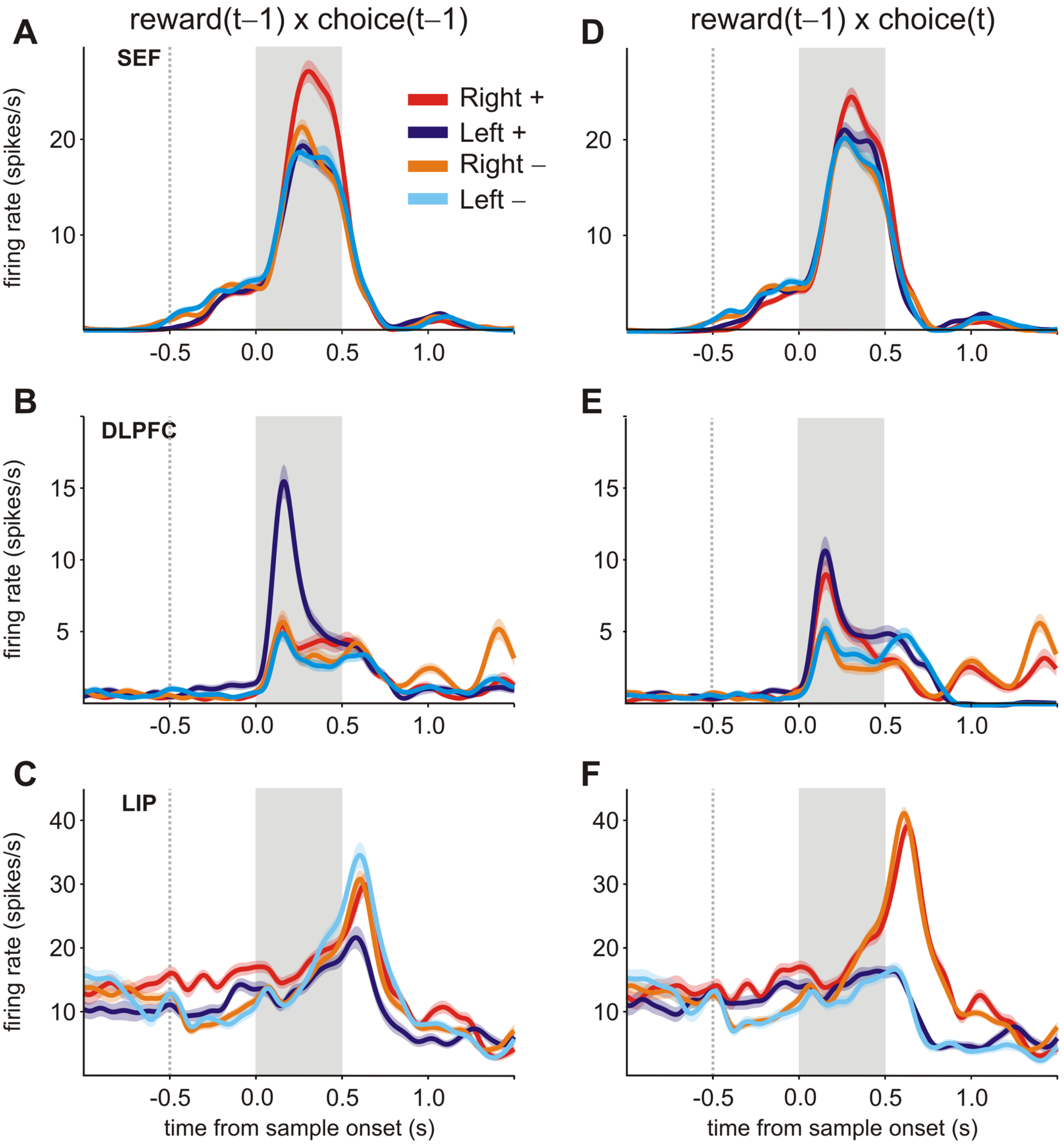

Figure 4.

Activity of example neurons from SEF, DLPFC, and LIP during the matching pennies task. **A–C**. Spike density functions (SDF) for single neurons for trials separated by the animal's choice (left or right) and outcome (rewarded, +, or non-rewarded, −) in the previous trial. For all 3 neurons (A, SEF; B, DLPFC; C, LIP), encoding of the previous choice was more robust when the previous trial was rewarded (2-way ANOVA, previous choice × previous reward interaction, $p<10^{-4}$). **D–F**. SDF for the same neurons for trials separated by the animal's choice in the current trial and the reward in the previous trial. The signals related to the upcoming choice was not significantly affected by the previous reward for any neuron

(2-way ANOVA, previous reward × current choice interaction, p>0.25). The shaded area indicates mean ± SEM.

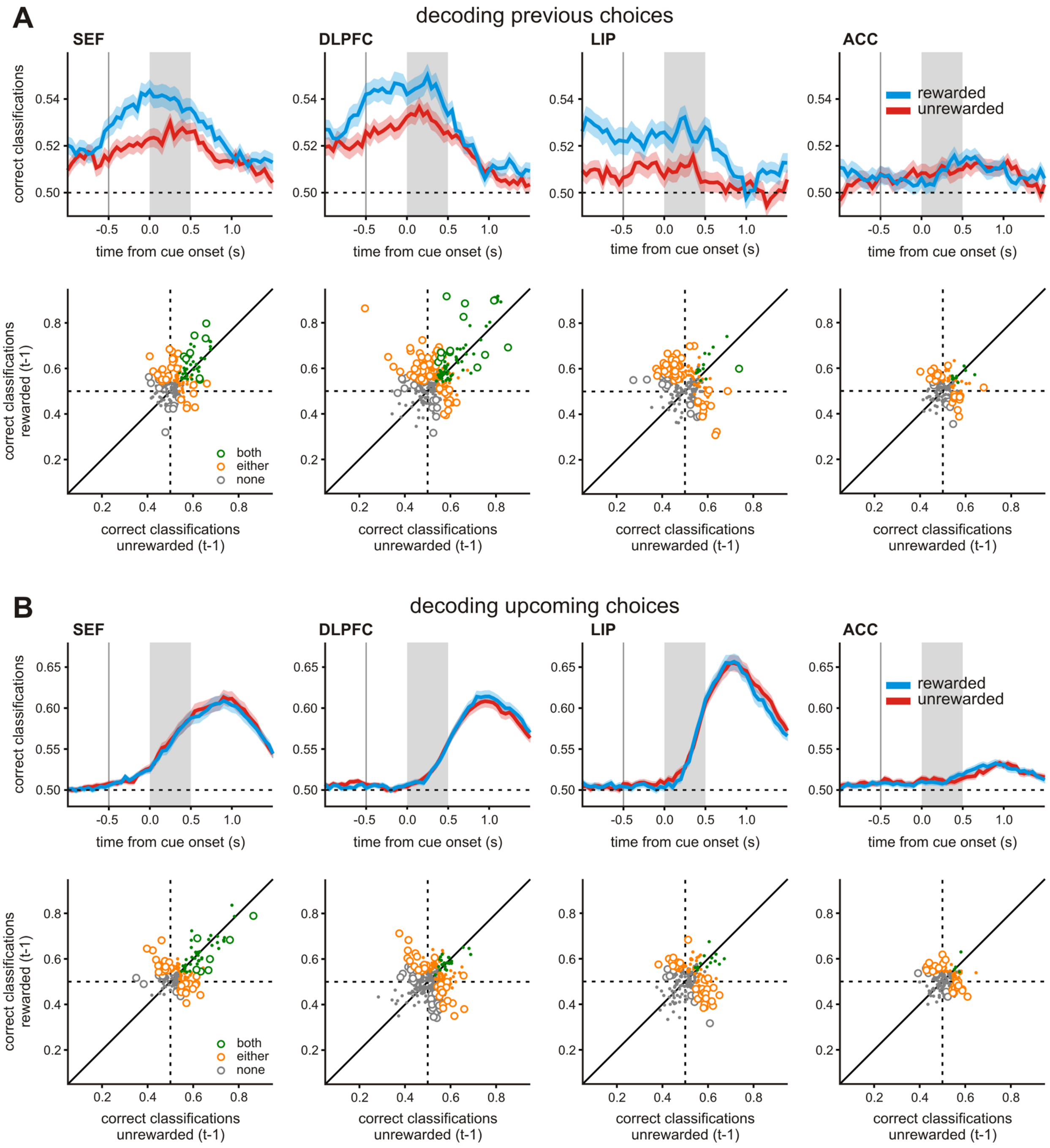

Figure 5.

Population summary for the effect of reward on the neural encoding of previous choice during the matching pennies task. **A**. Top. Average decoding accuracy for the previous choice estimated using a sliding window separately according to whether the animal's previous choice was rewarded or not. The shaded area indicates mean ± SEM. Bottom. Scatter plots show classification accuracy for each neuron following rewarded versus unrewarded trials. Symbol colors indicate whether the decoding accuracy was significantly above chance for both (green), either (orange), or neither (grey) of rewarded and unrewarded outcomes (z-test, p<0.05). Open and closed symbols indicate whether classification was significantly different for the two outcomes. **B**. The decoding accuracy for the animal's upcoming, shown in the same format as in **A**. See also Figures S2 and S3.

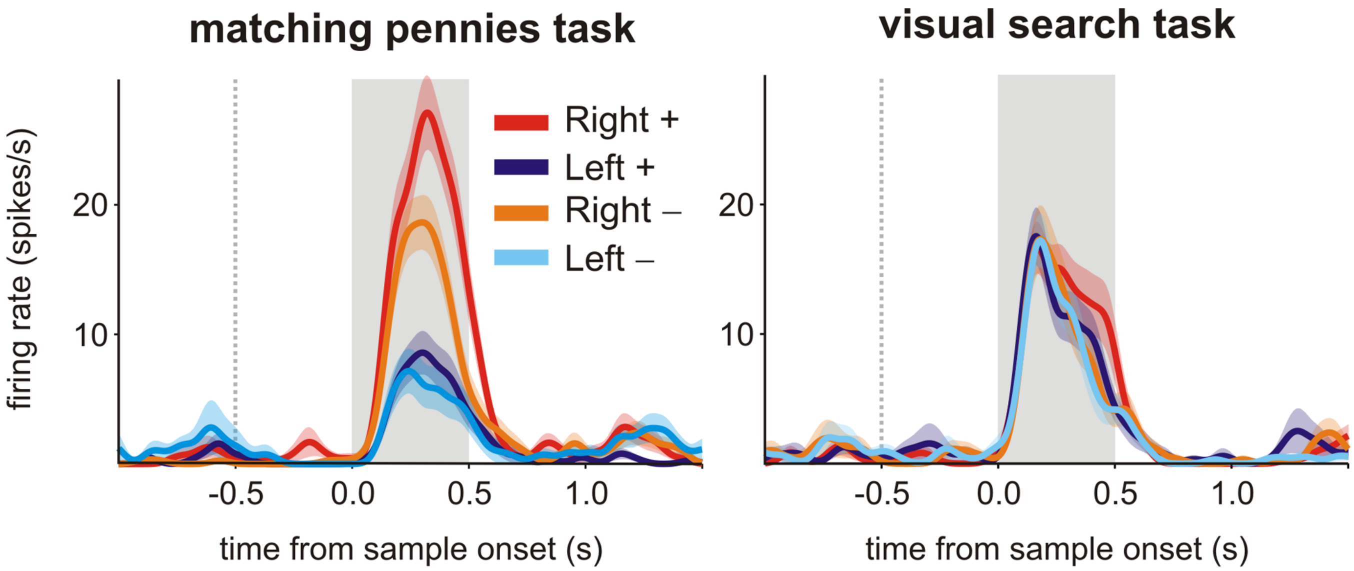

Figure 6.

Activity of an example SEF neuron during the visual search and matching pennies task. The number of trials used for calculate the spike density functions was equated for the two tasks (n=128). Same format as in Figure 4A–C.

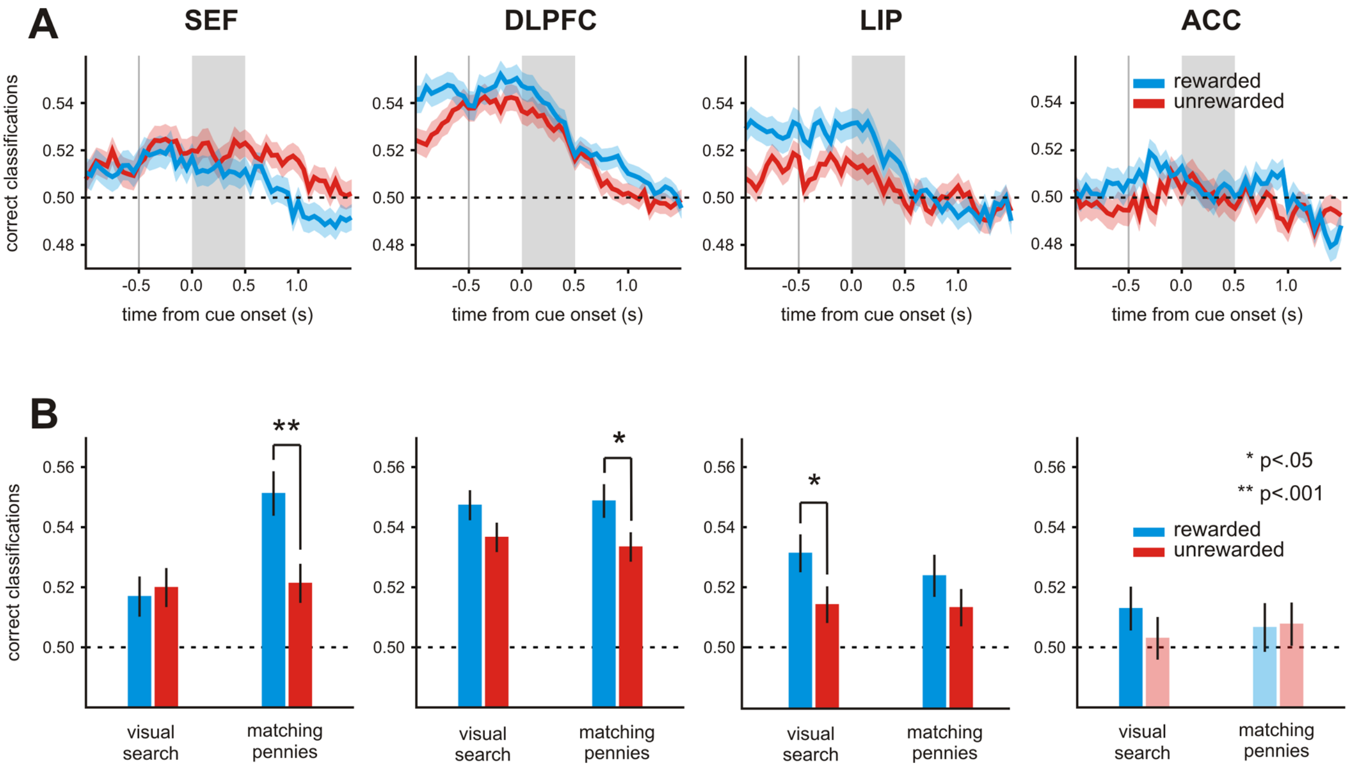

Figure 7.

Population summary for the task-specific effect of reward on the neural encoding of previous choice. **A**. Top. Decoding accuracy for the animal's previous choice in the visual search task. Same format as in Figure 5A. **B**. Average decoding accuracies for previous choice shown separately for each region, task, and previous reward. Saturated colors indicate that the accuracy was significantly higher than the chance level (t-test, p<0.05). Error bars indicate ± SEM. See also Figure S3.

Table 1

Summary of animal's choice behaviors. P(right), the probability of choosing the right target; P(reward), average reward rate; P(WSLS), the probability of using the win-stay-lose-switch strategy; P(WS)>P(LS), the percentage of sessions in which the probability of win-stay was larger than the probability of lose-switch.

Monkey (sessions)	P(Right)	P(reward)	P(WSLS)	P(WS)>P(LS)	RL>Equilibrium
C (n=21)	**48.7%** **(7, 33.3%)**	**47.6%** **(9, 42.9%)**	**53.6%** **(8, 38.1%)**	**76.2%** **(9, 42.9%)**	**(15, 71.4%)**
D (n=79)	**50.7%** **(13, 16.5%)**	**49.0%** **(11, 13.9%)**	**50.6%** **(25, 31.7%)**	**65.8%** **(30, 38.0%)**	(29, 36.7%)
E (n=111)	50.0% **(13, 11.7%)**	**48.8%** **(16, 14.4%)**	**51.2%** **(31, 27.9%)**	54.1% **(18, 16.2%)**	**(77, 69.4%)**
H (n=32)	49.8% **(7, 21.9%)**	**46.5%** **(7, 21.9%)**	**57.1%** **(23, 71.9%)**	59.4% **(13, 40.6%)**	**(25, 78.1%)**
I (n=47)	**49.1%** **(9, 19.2%)**	**47.9%** **(12, 25.5%)**	**51.4%** **(13, 27.7%)**	36.2% **(10, 21.3%)**	(29, 61.7%)
K (n=29)	**51.0%** **(7, 24.1%)**	**48.4%** **(5, 17.2%)**	**55.2%** **(12, 41.4%)**	**69.0%** **(12, 41.4%)**	(16, 55.2%)
All	50.0% **(56, 17.6%)**	**48.5%** **(94, 29.5%)**	**52.2%** **(112, 35.1%)**	**57.7%** **(92, 28.8%)**	**(191, 59.9%)**

For each of these 4 columns, the overall proportion of the trials or sessions are shown first, with the numbers inside parentheses corresponding to the number and percentage of sessions in which the probability was significantly different from 0.5 (two-tailed binomial tests, p<0.05). Bold typeface indicates that the results were significantly different from 0.5 (two-tailed binomial test, p<0.05) or higher than the chance level (one-tailed binomial test, p<0.05). RL>Equilibrium, the number and percentage of sessions in which the modified reinforcement learning (RL) model with separate win and loss reward parameters accounted for the animal's choice behavior better than the equilibrium strategy according to the Bayesian information criterion (BIC). Bold type face indicates that the proportion was significantly higher than 50% (one-tailed binomial test, p<0.05). See also Table S1.