Published in final edited form as:

Neuron. 2018 August 08; 99(3): 598–608.e4. doi:10.1016/j.neuron.2018.06.033.

Volatility facilitates value updating in the prefrontal cortex

Bart Massi[1,•], **Christopher H. Donahue**[2,•,†], and **Daeyeol Lee**[1,2,3,4,5,*]

[1]Interdeparmental Neuroscience Program, Yale School of Medicine, New Haven, CT 06510, USA

[2]Department of Neuroscience, Yale School of Medicine, New Haven, CT 06510, USA

[3]Department of Psychiatry, Yale School of Medicine, New Haven, CT 06510, USA

[4]Kavli Institute for Neuroscience, Yale School of Medicine, New Haven, CT 06510, USA

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

SUMMARY

Adaptation of learning and decision-making might depend on the regulation of activity in the prefrontal cortex. Here, we examined how volatility of reward probabilities influences learning and neural activity in the primate prefrontal cortex. We found that animals selected recently rewarded targets more often when reward probabilities of different options fluctuated across trials than when they were fixed. Additionally, neurons in the orbitofrontal cortex displayed more sustained activity related to the outcomes of their previous choices when reward probabilities changed over time. Such volatility also enhanced signals in the dorsolateral prefrontal cortex related to the current, but not the previous, location of the previously rewarded target. These results suggest that prefrontal activity related to choice and reward is dynamically regulated by the volatility of the environment, and underscore the role of the prefrontal cortex in identifying aspects of the environment that are responsible for previous outcomes and should be learned.

BLURB

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

[•]These authors contributed equally to this work.

[†]Present address: Gladstone Institute of Neurological Disease, San Francisco, CA 94158, USA.

Lead Author: Daeyeol Lee

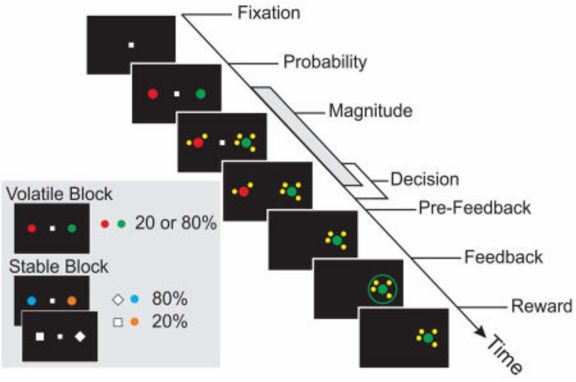
Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published 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.

AUTHOR CONTRIBUTIONS

C.H.D. and D.L. designed the experiments. C.H.D. collected the data. B.M. and C.H.D. analyzed the data, and all authors contributed to writing the manuscript.

COMPETITING FINANCIAL INTERESTS

The authors declare no competing interests.



Massi et al. show that signals in the prefrontal cortex related to choices and outcomes are enhanced when reward probabilities are volatile rather than stable. Furthermore, when reward probabilities are volatile, rewards strengthen task-relevant, but not task-irrelevant, signals.

INTRODUCTION

What and how an animal should learn depends on the stability of the animal's environment. For example, even when outcomes of actions chosen by decision makers are stochastic and unpredictable, if the probabilities of different outcomes from actions are fixed and precisely known, it is not necessary to update decision-making strategies. By contrast, if little is known about the probabilities of different outcomes, the outcome from each action is likely to have stronger influence on the decision-maker's subsequent strategies. Previous studies have shown that humans and other animals adjust their learning strategies nearly optimally depending on the level of uncertainty or volatility of their environment (Behrens et al., 2007; Nassar et al., 2012; Lee et al., 2014; McGuire et al., 2014). In addition, given the role of the prefrontal cortex (PFC) in learning and decision making (Miller and Cohen, 2001; Wallis and Kennerley, 2010), flexible regulation of activity in the PFC might facilitate different learning strategies (Doya, 2002; Yu and Dayan, 2005; Nassar et al., 2012; Tervo et al., 2014; Farashahi et al., 2017). In particular, recurrent connections in PFC networks are thought to facilitate persistent activity (Compte et al., 2000; Wang et al., 2013; Murray et al., 2014; Chaudhuri et al., 2015), and relatively small changes in the strength of recurrent connections determines whether neural circuits respond to stimuli from the recent or distant past (Chaudhuri et al., 2015; Wong and Wang, 2006; Murray et al., 2012). Persistent PFC signals related to an animal's previous choices and their outcomes might contribute to multiple aspects of reinforcement learning, such as value updating and temporal credit assignment (Curtis and Lee, 2010; Walton et al., 2010; Bernacchia et al., 2011; Donahue et al., 2013; Donahue and Lee, 2015; Asaad et al., 2017). However, how PFC signals related to an animal's choices and their outcomes might be modulated by the stability or volatility of the environment is not known.

To address this question, we trained rhesus monkeys to perform a probabilistic reversal learning task in which we manipulated the volatility of the reward probabilities associated with the available options. During volatile blocks, the reward probabilities of each option underwent periodic reversals (Donahue and Lee, 2015), which encouraged the animals to choose their actions according to their recent choice and reward histories. By contrast, during stable blocks, the reward probability for each target was fixed throughout the experiment. We found that past outcomes tended to influence the animal’s behavior strongly during the volatile blocks but not the stable blocks. In addition, single neurons in the prefrontal cortex tended to have a stronger representation of past choices and outcomes during the volatile blocks than during the stable blocks. Previous rewards also enhanced the neural signals that combined information about the events in the previous trial with stimulus parameters of the current trial, but only during the volatile blocks. These findings suggest that the representation of reward in the PFC depends on environmental volatility and shapes the encoding of other task-relevant information for learning.

RESULTS

Volatility Promotes Learning

To test whether volatility affects the nature of PFC signals that are important for decision making, we trained two monkeys (U and X) to perform a probabilistic reversal task (Figure 1A). In each trial, the animal was required to select one of two peripheral targets with variable reward probabilities and magnitudes. In a given block of trials, one target color or shape was associated with a high reward probability (80%), and the other with a low reward probability (20%). During the volatile blocks, target colors associated with high and low reward probabilities were switched after 20 or 40 trials (Donahue and Lee, 2015), whereas during the stable blocks, target colors or shapes corresponding to high and low reward probabilities were fixed throughout the entire experiment. As expected, the animals adjusted their preference for different target colors based on the outcomes of their recent choices significantly and substantially more during the volatile blocks than during the stable blocks (Figures 2A and S1A). During the volatile blocks, the animals rapidly adjusted their preferences following reversals regardless of the length of the preceding volatile sub-block (Figure S1B). Learning rates estimated with a reinforcement learning model (Sutton and Barto, 1998) were significantly larger (paired t-test, $p<10^{-59}$ for both animals) during the volatile blocks (both animals $\alpha_v = 0.22$; Figure 2B) than during the stable blocks ($\alpha_s = 0.01$ and 0.02 for monkeys U and X, respectively). We also tested a model with two separate learning rates for rewarded and unrewarded outcomes and found that the learning rate varied with both volatility (main effect of volatility in a 2-way ANOVA, $p<10^{-16}$ for both animals) and outcome (reward main effect, $p<10^{-16}$ for both animals). In addition, the effect of reward on learning rates (mean±s.d.) was significantly larger for the volatile block than for the stable block ($\alpha_{v\text{-rewarded}} = 0.39\pm 0.13$, $\alpha_{v\text{-unrewarded}} = 0.22\pm0.07$, $\alpha_{s\text{-rewarded}} = 0.13\pm0.12$, $\alpha_{s\text{-unrewarded}} = 0.00\pm 0.01$; reward × volatility interaction, $p<0.005$ for both animals). The relative reward magnitude also influenced behavior. The regression coefficient (mean±s.d.) for the difference in magnitude between the two options (see STAR Methods) in our reinforcement learning model was significantly above zero for both monkeys (1.35± 0.31 and 1.39±0.21, for monkeys U and X, respectively; one-sample t-tests, $p<10^{-111}$ for both animals). Thus, animals’ decisions depended both on their estimates of reward probabilities and the explicit cues associated with the reward magnitude for each option.

Multiplexing of Task-related Signals in the Prefrontal Cortex

While the animals performed the task, we recorded single-neuron activity from three regions in the PFC (N=174 neurons in dorsolateral prefrontal cortex, DLPFC; N=135 in orbitofrontal cortex, OFC; N=135 in anterior cingulate cortex, ACC; Figure 1B) to test whether task-related activity was influenced by volatility of reward probabilities or the outcomes of previous trials. We focused on how different types of task-related information included in neural activity was modulated by reward and learning (Donahue et al., 2013; Donahue and Lee, 2015; Histed et al., 2009). Consistent with previous findings (Donahue et al., 2013; Kennerley et al., 2011; Cai and Padoa-Schioppa, 2014), we identified several types of signals related to the visual stimuli and actions in the activity recorded from three cortical areas. Some of these signals were not affected by volatility of reward probabilities or the outcomes of previous trials. To examine the strength of signals and how they were affected by volatility, we calculated the coefficient of partial determination (CPD; Kim et al., 2008), or proportion of variance accounted for by variables of interest. To examine more closely how the outcome of the previous trial impacted neural representation of various factors, we further examined these signals using a decoding analysis (see STAR Methods).

During the interval immediately after target onset (post-target period; see STAR Methods), the DLPFC tended to encode the signals related to the animal’s previous action more strongly than those in the OFC and ACC, and similarly for volatile and stable blocks (main effect of region in a region × volatility two-way mixed ANOVA on CPD, $p<10^{-3}$; main effect of region in a region × volatility × reward 3-way mixed ANOVA on decoding accuracy, $p<10^{-4}$; Figures 3A, 3B, and S2). Consistent with the findings in our previous study (Donahue and Lee, 2015), we did not find any evidence that the decoding accuracy for the animal’s action in the previous trial was affected by reward (main effect of reward and reward × volatility interaction in a repeated measures ANOVA, $p > 0.05$ for all cortical areas; Figures 3C, 3D, and S2B). By contrast, we had also shown that during the matching pennies task, signals related to previous action are enhanced in the DLPFC when the animal was rewarded in the previous trial (Donahue et al., 2013). We confirmed that the difference between accuracy of decoding the previous action for rewarded and unrewarded trials during the volatile block for the DLPFC in the present study was indeed significantly smaller than the corresponding effect previously observed during the matching pennies task (Donahue et al., 2013; one-sample t-test, $p<10^{-3}$). Given that animals learned the values of target locations or actions during the matching pennies task, but not during the tasks used in the present study, this suggests that representation in the DLPFC might be modulated by reward only when it is behaviorally relevant (Donahue et al., 2013; Donahue and Lee, 2015).

Like the signals related to previous choice, signals related to the animal’s upcoming choice in the DLPFC were stronger than those in the OFC and ACC (main effect of region in a region × volatility two-way repeated measures ANOVA on CPD, $p<0.05$; Figures 4A, 4B, and S3), but did not differ for the volatile and stable blocks (paired t-test on CPD, $p=0.07$). By contrast, signals related to the upcoming choice were unreliable in the OFC (one-sample t-tests, $p>0.1$ for rewarded and unrewarded trials in both volatile and stable blocks; Figures 4C and 4D), whereas they were weakly but significantly present in the ACC (one-sample t-tests, $p<0.05$ in all cases; Figures 4C, 4D, and S3B). The accuracy of decoding upcoming choice was not affected by the outcome of the previous trial (main effect of reward in a reward × volatility 2-way repeated measures ANOVA, $p>0.05$) in any of three cortical areas.

During the task used in this study, a target's color or shape consistently signaled the reward probability of that target during the stable blocks, but the reward probability of a given target color varied during the volatile blocks. Therefore, information about the identity of the target in a given location should be more reliably decoded in the stable condition from the activity of neurons encoding the position of the high-reward probability target. Indeed, the color or shape of the target in a given location could be reliably decoded for the DLPFC neurons in the stable condition, but not in the volatile condition (main effect of volatility in a reward × volatility 2-way repeated measures ANOVA, $p<0.001$; Figures 5 and S4). By contrast, the identity of the target in a given location could not be decoded reliably in the OFC, whereas this was weakly represented in the ACC only after unrewarded trials (one-sample t-test, $p<0.05$). More importantly, the accuracy of decoding the target identity was not significantly different for volatile and stable conditions in the OFC or ACC ($p>0.1$ in both cases). This difference across 3 cortical areas was statistically significant (region × volatility interaction in a region × volatility × reward 3-way mixed-ANOVA, $p<0.01$). However, the accuracy of decoding the target identity was not consistently affected by the outcome of the previous trial in any cortical areas (main effect of reward and reward × volatility interaction in a repeated measures ANOVA, $p > 0.05$ for all cortical areas).

Volatility Enhances the Prefrontal Activity related to Past Events

In contrast to the signals related to the actions chosen in the previous and current trials, we found that the signals relevant for learning were encoded more robustly during the volatile blocks than during the stable blocks. In particular, the color of the target chosen by the animal in the previous trial and the outcome of this choice were key for learning for the task used in the present study, and activity in the PFC related to these two variables were represented more strongly during the volatile blocks.

We found that across all cortical areas in aggregate, target color chosen in the previous trial was reliably decoded, but only when the animal was rewarded in the previous trial during the volatile block (volatility × reward interaction in volatility × reward × region 3-way mixed ANOVA, $p<0.05$; Figures 6 and S5). For example, during the post-target period, the DLPFC represented the target color chosen in the previous trial more accurately when the animal was rewarded in the previous trial during the volatile blocks than when it was not rewarded (paired t-test on decoding accuracy, $p<0.005$). The effect of previous reward was not significant for the stable blocks ($p=0.7$), although this difference between the volatile and stable blocks was not statistically significant in any individual cortical region (reward × volatility interaction in 2-way repeated measures ANOVA, $p>0.1$ for all areas). The effect of volatility on the neural activity related to the previously chosen color was similar in the OFC (Figure 6 and S5). In particular, the outcome of the previous trial impacted the representation of the previously chosen target color in the OFC (main effect of reward in a 2-way repeated measures ANOVA, $p<0.005$). During the volatile blocks, target color chosen in the previous trial was decoded more accurately from the activity of OFC neurons when the animal was rewarded in the previous trial than when it was not rewarded (paired t-test,

$p<0.005$). However, during the stable blocks, this difference was not significant ($p=0.41$). For ACC, the accuracy of decoding previously chosen color was not significantly influenced by reward, volatility, or their interaction in the ACC ($p>0.3$ in all cases), although previously chosen color could be decoded reliably when the animal was rewarded in the previous trial during the volatile blocks (Figure 6C).

We also found that in the OFC, neural signals related to the previous outcome were more robust during the volatile blocks than during the stable blocks. For example, during the 0.5 s interval before target onset, the activity of the OFC neuron shown in Figure 7A was modulated by the previous outcome more strongly in the volatile blocks than in the stable blocks (n=585 trials, volatility × reward interaction in 2-way ANOVA, $p<0.05$). During the same period, a significant fraction of neurons in the OFC (N=14 neurons, 10.4%; binomial test, $p<0.01$) modulated their activity more strongly according to the previous outcome in the volatile blocks than in stable blocks, whereas the percentage of such neurons in the DLPFC (N=9 neurons, 5.2%) and ACC (N=6 neurons, 4.4%) was not significantly above chance ($p>0.4$ for both areas). For the OFC, we also found that the effect size (CPD) of the activity modulation related to the previous outcome was significantly larger in the volatile blocks than in the stable blocks (paired t-test, $p=0.03$; Figure 7B). This difference was not significant for the ACC ($p=0.06$) or DLPFC ($p=0.65$). We also analyzed the signals related to the feedback in the current trial during a 1-s window starting 0.25 s after feedback onset and found that the outcome of the current trial was represented by many neurons in the DLPFC (N=108 neurons, 62.1%), the OFC (N=76, 56.3%), and the ACC (N=91, 67.4%). In addition, we found that the mean CPD for the current outcome was significantly stronger in the ACC (main effect of region in a region × volatility mixed effects ANOVA, $p<0.01$). Nevertheless, we found no evidence that this information was affected by volatility (main effect of volatility, $p>0.5$; Figure S6A). In addition to information about the choice and outcome of the previous trial, information about reward magnitude was represented in the PFC. For example, a significant fraction of neurons in the DLPFC (N=24, 13.8%; binomial test, $p<10^{-5}$), OFC (N=28, 20.7%; binomial test, $p<10^{-9}$), and ACC (N=30, 22.2%; binomial test, $p<10^{-11}$) represented the difference between chosen and unchosen reward magnitudes. The strength of this magnitude information varied significantly across different cortical areas (effect of region in region × volatility mixed effects ANOVA, $p<0.01$), and was greater in the ACC than in the OFC ($p<0.004$). By contrast, volatility did not significantly affect the amount of magnitude-related information at the population level in any cortical area (Figure S6B).

Volatility Facilitates Updating of Value Signals in the PFC

We have hypothesized that the PFC might be involved in estimating the position of the target associated with a higher reward probability by combining the relevant information about the animal's previous experience and incoming sensory stimuli (Donahue and Lee, 2015). During the volatile condition of the task used in the present study, this can be accomplished by integrating 3 different types of information: the target color chosen in the previous trial, the outcome of that choice, and the current position of the same target color. For example, if the animal was rewarded for choosing a red target in the previous trial, and if the same red target appears on the right side in the current trial, then the rightward target is more likely to have a higher reward probability in the current trial. The rightward target would be also deemed desirable by the animal when the previously chosen but unrewarded target appears on the left. In the following, the target color favored by the animal's choice and its outcome in the previous trial would be referred to as the high-value target, and its current location (i.e., current position of the previous chosen color × previous reward interaction) as the high-value location (HVL).

We found that signals related to HVL were represented most strongly in the DLPFC and were stronger in the volatile blocks than the stable blocks. For example, during the volatile blocks, the activity of the DLPFC neuron illustrated in Figure 8A was higher during the 0.5 s period beginning 0.25 s after target onset when the HVL was left than when it was right. This difference was significantly attenuated for the stable blocks (n = 640 trials, current position of the previous chosen color × previous reward × volatility interaction in 3-way ANOVA, $p<0.05$). During the same epoch, the proportion of neurons with activity related to the HVL that was significantly modulated by volatility was significantly above the chance level for the DLPFC (N=20 neurons, 11.5%; binomial test, $p<10^{-3}$), but not for the OFC (N=11, 8.1%; binomial test, $p=0.08$) or the ACC (N=10, 7.4%; binomial test, $p=0.14$). A significant fraction of DLPFC neurons displayed stronger signals related to the HVL during the volatile blocks than during the stable blocks in both monkey U (11/102 neurons, 10.8%; binomial test, $p<0.05$) and monkey X (9/72 neurons, 12.5%; binomial test, $p<0.01$). Although this difference in the overall proportion of neurons across cortical areas was not statistically significant (χ^2-test, $p>0.4$), analysis of the effect size (CPD) and decoding accuracy indicated that volatility influenced the activity related to HVL in the DLPFC. For example, during the post-target period, the average CPD for the HVL was significantly larger during the volatile blocks than during the stable blocks (main effect of volatility in a region × volatility 2-way mixed ANOVA, $p<0.01$). This effect was significant only in the DLPFC (paired t-test, $p<0.0005$), and significantly weaker in other cortical areas (region × volatility interaction, $p<0.005$; Figure 8B). Similarly, both previous reward and volatility affected the accuracy of decoding the current position of the previously chosen target color only in the DLPFC (region × volatility, region × reward, and reward × volatility interactions in a reward × region × volatility 3-way mixed ANOVA, $p<0.05$ in all cases; Figures 8C, 8D, and S7). The average decoding accuracy for the current position of the previously chosen color was significantly higher after rewarded trials than after unrewarded trials in the DLPFC during the volatile blocks, but this difference was not significant during the stable blocks (previous reward × volatility interaction in a 2-way repeated measures ANOVA, $p<0.005$; Figures 8C, 8D, and S7B). In addition, the difference in the decoding accuracy for the current position of previously rewarded and unrewarded target color was stable throughout the volatile block, as it did not differ significantly between the first and subsequent sub-blocks in the volatile condition (paired t-test, $p>0.5$). By contrast, the signals related to the HVL were largely absent in the OFC and ACC (Figures 8C and 8D). These results suggest that the DLPFC might play a special role in estimating the desirable target location according to the animal's previous experience in volatile environments.

In contrast to signals related to the position of the previously rewarded target color in the volatile block, signals related to the position of the high reward-probability target in the stable block were unaffected by events in the previous trial (Figure 5). To further test whether the previous outcome enhanced the signals related to the HVL selectively in the volatile condition, we tested whether the signals related to the HVL in the volatile condition and the position of the high reward-probability target in the stable block were similarly affected by the previous outcome. To do so, we examined decoding accuracy for the current position of the previously chosen target color in the volatile block following rewarded and unrewarded trials, and compared it to decoding accuracy for the position of the high vs. low reward-probability targets in the stable block following rewarded and unrewarded trials. We found that the previous outcome enhanced the accuracy of decoding HVL in the DLPFC during the volatile blocks significantly more than the signals related to the position of the high-reward probability target during the stable blocks (previous reward × volatility interaction in a 2-way repeated measures ANOVA, $p<10^{-4}$). Nevertheless, the same population of neurons in the DLPFC tended to represent both the position of the high reward probability target in the stable block and the position of the previously rewarded target color in the volatile blocks (correlation of regression coefficients, $r=0.42$, $p<10^{-7}$). Thus, our findings suggest that the same prefrontal cortical circuits might be flexibly tuned to compute values over different time scales in volatile and stable environments.

Finally, we directly tested the hypothesis that volatility causes reward to enhance the representation of task-relevant information, but not task-irrelevant information. To do this, we compared the neural representations of previous action (Figure 3) and the location of previously chosen target color (Figure 8) using a 3-way ANOVA on decoding accuracy in the DLPFC, with volatility, the outcome of the previous trial, and task-relevance (i.e., previous action vs. location of the previously chosen target) as factors. In this analysis, the three-way interaction reflects the degree to which task-relevance changes the effect of volatility on how reward enhances information representation. Significant interactions were found between previous outcome and volatility ($p<0.05$) as well as previous outcome and task relevance ($p<0.005$), although 3-way interaction was only marginally significant ($p=0.069$). Collectively, these results offer evidence that reward modulates task-relevant variables, but not task-irrelevant variables, when the environment is volatile.

DISCUSSION

During the task used in the present study, the animals adjusted their learning strategies according to environmental volatility. The animals assigned a higher weight to the outcomes of their recent choices, resulting in a higher learning rate when reward probabilities for alternative options were volatile, compared to when they were stable. This was paralleled by more robust encoding of reward signals in the prefrontal cortex, especially in the OFC, indicating that neural representations of recent outcomes are enhanced in volatile environments during learning. In addition, high volatility of reward probabilities altered how other task-relevant signals in the prefrontal cortex were affected by the outcomes of the animal’s previous choices. In particular, signals related to the current position of the previously chosen target were enhanced by reward in the DLPFC during the volatile blocks, but not during the stable blocks. By contrast, signals related to previous or upcoming actions were not influenced by volatility or previous outcome. Collectively, our results suggest that signals related to task-relevant variables in the PFC are flexibly combined with reward information to facilitate learning.

Volatility Modulates Choice and Outcome Signals

The results from this study suggest that prefrontal signals related to the outcome of previous choices might be selectively enhanced when they are useful for updating the values of prospective actions. Specifically, neural representations of the previous trial's outcome were strengthened during the volatile blocks in the OFC, which carries signals related to values and expected outcomes that are critical for decision making (Kennerley et al., 2011; Padoa-Schioppa and Assad, 2006). The OFC is thought to play causal roles in behavioral adjustment during decision-making tasks (Schoenbaum et al., 2002; Fellows and Farah, 2003; Rudebeck et al., 2008; Camille et al., 2011), so modulation of reward signals may be critical for adjusting stimulus-reward associations in volatile environments. While some studies have shown that simply learning the best option during a reversal learning task (i.e., model-free reinforcement learning; Sutton and Barto 1998) does not depend on the OFC (Kazama and Bachevalier, 2009; Rudebeck et al., 2013), the finding that reward-related signaling in the OFC changes with environmental volatility is consistent with the previously identified role of this region in model-based reinforcement learning (Jones et al., 2012) and the representation of behavioral strategy (Chau et al., 2015). In particular, the results in the present study suggest that environmental volatility might alter the timescale of reward representation in the OFC, which may facilitate reinforcement learning (Bernacchia et al., 2011; Murray et al., 2014; Meder et al., 2017).

In addition to the outcome of previous choices, information about the previously chosen target color was critical for making valuable choices during the volatile blocks of the task used in this study. We found that neurons in the DLPFC and OFC tended to exhibit a stronger representation of the previously chosen target color following rewarded trials during the volatile blocks. By contrast, prefrontal signals related to the previously chosen target color were weaker and unaffected by prior outcomes during the stable block. These results add to the evidence that DLPFC signals related to task-related variables, such as the animal's choice in the previous trial, might be enhanced by the previous outcome only when they are relevant for learning. For example, several studies have found that neural signals related to the animal's previous or current choice in the prefrontal cortex and striatum are enhanced by the outcome of the animal's previous choice (Donahue et al., 2013; Histed et al., 2009; Ito and Doya, 2009). Consistent with the results from a previous study (Donahue and Lee, 2015), however, we found that when the animal's previous action was not relevant for decision-making in the current trial, signals related to the previous action were still encoded in the DLPFC but not affected by the previous outcome.

Previous research has shown that neurons in the ACC carry signals related to learning, reward, and decision-making (Seo and Lee, 2007; Hayden et al., 2009; Hayden and Platt, 2010; Kennerley et al., 2011). Furthermore, neuroimaging work has shown that volatility may influence learning-related signals in this region. The BOLD signals in the ACC track environmental volatility (Behrens et al., 2007), and the learning rate for the value representation in the ACC increases in volatile environments compared to stable environments (Meder et al., 2017). We did not find robust learning-related signals that were influenced by volatility in the ACC at the single neuron level, and this might be because volatility was explicitly indicated by sensory features of the targets in the present study. We found that many neurons in the ACC carried signals related to the relative reward available from chosen and unchosen targets. During the task used in this study, the animals had to combine learned reward probabilities with information about reward magnitude, and therefore neural signals in the ACC related to the relative reward available from both options might contribute to reward-maximizing behavior. Moreover, the ACC had stronger representation of the signals related to the outcome in the current trial. These results are consistent with those of previous studies implicating ACC in flexible decision-making (Heilbronner and Hayden, 2016).

Prefrontal Contributions to Flexible Value Updating

The results from this study indicate that in the DLPFC reward signals were flexibly combined with other mnemonic and sensory signals relevant to updating the animal's behavioral strategies. In particular, signals related to the target color chosen in the previous trial and the current location of the previously chosen color in the DLPFC were enhanced by previous outcome only during the volatile block. Consistent with the results from previous studies (Cai and Padoa-Schioppa, 2014; Kim et al., 2012), these results suggest that the DLPFC plays an important role in transforming multiple types of signals relevant for decision making in action frame of reference, and that heterogeneous conjunctive codes might be essential for this transformation (Fusi et al., 2016). Reward information may be dynamically routed to the DLPFC in volatile environments (Donahue and Lee, 2015), perhaps via the ACC and OFC, to support the formation of these conjunctive signals. Moreover, neurons that represented information about the current location of the previously rewarded target color in the volatile block also tended to represent the position of a high reward-probability target in the stable block. Therefore, DLPFC neurons might reflect the values of specific choices estimated on different time scales according to the volatility of the environment.

Previous work has indicated that prefrontal signals related to choice and outcome are represented in multiple time scales (Bernacchia et al., 2011; Meder et al., 2017). In this study, we showed that the strength of such persistent activity and its effect on other types of prefrontal signals is modulated by the volatility of reward probabilities. Given that timescale of persistent activity can be controlled by recurrent excitation (Wang et al., 2008; Murray et al., 2014; Chaudhuri et al., 2015), volatility of the animal's environment might influence the neural activity related to learning and decision making by controlling the efficacy of local recurrent connections in the PFC. Alternatively, the effects of volatility on prefrontal signals might be mediated by short-term plasticity induced by dopamine receptor signaling (Seamans et al., 2001; Young and Yang, 2005; Soltani et al., 2013). Currently, the precise underlying neural mechanism for altering the time scale of integration in the prefrontal cortex remains unknown (Farashahi et al., 2017). Nevertheless, prefrontal signals modulated by volatility and previous outcomes identified in the present study might contribute to the identification of the features relevant for reinforcement learning and hence the resolution of the temporal credit assignment problem (Sutton and Barto, 1998).

STAR METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Daeyeol Lee (daeyeol.lee@yale.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Two healthy male rhesus monkeys (U and X) were used. They weighed approximately 13 and 9 kg, and were 6 and 7 years old, respectively, during the experiment. Monkey X was experimentally naïve, whereas monkey U was used for a previous experiment utilizing a similar task (Donahue and Lee, 2015). Eye movements were monitored at a sampling rate of 225 Hz with an infrared eye tracker (ET49, Thomas Recording, Germany). All procedures were approved by the Institutional Animal Care and Use Committee (IACUC) at Yale University.

METHOD DETAILS

Behavioral Task—To assess how animals integrate choice and outcome history under varying levels of environmental volatility, we trained the animals to perform a probabilistic reversal task where the volatility of target reward probabilities was manipulated across blocks (Figure 1A). To initiate a trial, the animals were required to maintain fixation on a small white square (0.7° × 0.7°) in the center of a computer monitor for 0.5 s. Next, two peripheral targets were presented along the horizontal meridian (diameter=1.4°). In a given block of trials, one of the targets was associated with a high reward probability (80%) and the other was associated with a low reward probability (20%). After a 0.5 s interval (target period), a set of small yellow tokens (diameter=0.6°) were presented around each target indicating the magnitude of available reward. The central fixation cue was extinguished after a random interval ranging from 0.5 to 1.2 s following magnitude onset according to a truncated exponential distribution (min = 500 ms, mean = 705 ms). After the central fixation target disappeared, the animals were free to shift their gaze towards one of the two peripheral targets. Following fixation on the chosen target for an additional 0.5 s (pre-feedback period), the animals received visual feedback indicating the trial's outcome. The feedback was a colored ring shown for 0.5 s around the chosen target (red or green in rewarded trials; blue or gray in unrewarded trials). In rewarded trials, the animals received the magnitude of juice according to the number of tokens (0.1 ml/token) after the after offset of the feedback rings. The inter-trial interval was 1 s.

The volatility of target reward probabilities was indicated by the visual characteristics of the targets in each block. In volatile blocks, the animals were presented with a red and green disks. One of the colors was associated with an 80% reward probability and the other was associated with a 20% reward probability. The identity of the high reward probability target underwent periodic reversals (20 or 40 trials) so that the animals had to estimate the reward probability for each target from their recent experience. In stable blocks, the animals were presented with either a pair of colored targets or shape targets, which were presented in smaller sub-blocks (20 to 40 trials). During color sub-blocks, the two targets were orange and cyan disks, respectively, whereas and during shape sub-blocks, they were a white diamond and a white square (Figure 1A, inset). The reward probability associated with each target in the stable blocks was fixed for the entire course of the experiment (across all sessions) such that one of the color (shape) targets was always associated with an 80% reward probability and the other color (shape) target was always associated with a 20% reward probability. The identity of the high reward probability target was counterbalanced across monkeys so that the cyan disk and white diamond represented the high reward probability target in monkey U and the low reward probability targets in monkey X, respectively.

Volatile and stable blocks alternated every 80 trials. In volatile blocks, the reward probabilities associated with the two targets remained fixed for sub-blocks with a length of 20 or 40 trials. To ensure that the identity of the high reward probability target was evenly distributed between the red and green targets, the reversal pattern was randomly drawn from the following 3 sequences of trial numbers: 20-20-20-20, 20-40-20, and 40-40. In stable blocks, transitions between sub-blocks in the color and shape condition underwent the same trial structure above, assuring that the animals sampled the color and shape sub-blocks evenly. The magnitudes of rewards associated with the two targets were drawn from the following 10 possible combinations: ({1,1} {1,2}, {1,4}, {1,8}, {2,1}, {2,4}, {4,1}, {4,2}, {4,4}, {8,1}). Each magnitude pair was counter-balanced across target locations, yielding 20 unique trial conditions that were presented in a pseudo-random order. A custom-developed Windows software (Picto) was used to deliver all visual stimuli and control the experiment.

Neurophysiological Recording—Single neuron activity was recorded from the DLPFC, OFC, and ACC of both monkeys (Figure 1B) using a 5-channel or 16-channel multielectrode recording system (Thomas Recording, Germany) and a multichannel acquisition processor (Tucker-Davis Technologies, FL or Plexon, TX). For the recordings in DLPFC, the recording chamber was centered over the left principal sulcus, and was located 9 mm (monkey X) and 10 mm (monkey U) anterior to the genu of the arcuate sulcus based on magnetic resonance images. Area 13 in OFC was targeted using the same recording chamber used for DLPFC. A cannula was used to guide the electrodes for these recordings. Neurons in ACC were recorded from the dorsal bank of the cingulate sulcus (area 24c; Seo and Lee, 2007).

All neurons in the dataset were recorded for a minimum of 320 trials (mean = 518.3 trials; standard deviation = 162.0 trials). The dataset consisted of 174 neurons in DLPFC (102 in monkey U, 72 in monkey X), 135 neurons in OFC (76 in monkey U, 59 in monkey X), and 135 neurons in ACC (72 in monkey U, 63 in monkey X). Data were collected during 182 sessions in monkey U, and 134 sessions in monkey X. Neurons were not pre-screened prior to collection, and all well-isolated neurons were recorded using Plexon or Tucker-Davis online spike-sorting software and included in subsequent analysis.

QUANTIFICATION AND STATISTICAL ANALYSIS

Analysis of Behavioral Data—All analysis of behavioral data was conducted using custom software written in MATLAB (Mathworks, MA). To perform the task used in this experiment, the animal would be expected to combine the reward magnitude and reward probability for each target. The magnitude information was given explicitly by the visual cues, whereas the probability information needed to be estimated through experience. Furthermore, the estimate for reward probabilities should be updated more frequently during volatile blocks than during stable blocks. To investigate how animals estimated reward probabilities in each block type, several variations of reinforcement learning models (Sutton and Barto, 1998) were fit to the animal's choice data.

To explicitly test whether and how volatility affected learning, we examined a model with two separate learning rates for volatile and stable blocks. Following each trial's outcome, estimates of the reward probability from the chosen target were updated using the following equation:

$$P_c(t+1) = P_c(t) + \alpha[R(t) - P_c(t)],$$

where $P_C(t)$ is the estimate of the reward probability for the target chosen in trial t, $R(t)$ is the outcome in trial t (1 if rewarded, 0 otherwise), and α is the learning rate. We fit separate learning rates in volatile (α_v) and stable (α_s) blocks. We also tested a model in which learning rate varied separately for rewarded and unrewarded outcomes in both volatile and stable blocks (Donahue and Lee, 2015). Reward probabilities were combined with magnitudes additively, according to the following logistic regression model, in order to compute the likelihood of a choice on each trial, as we have found this model to perform better than a multiplicative model (data not shown), consistent with the results our previous study (Donahue and Lee, 2015).

$$\text{logit } p(\text{right}) = \beta_0 + \beta_1 \Delta P_{LR}(t) + \beta_2 \Delta M_{LR}(t),$$

where p(right) is the probability of choosing the target on the right, $\Delta P_{LR}(t) = P_{right}(t) - P_{left}(t)$ is the difference between the estimated reward probabilities of the targets on the right and left in trial t, $\Delta M_{LR}(t) = M_{right}(t) - M_{left}(t)$ is the difference between the magnitudes of the targets on the right and left in trial t, and β_0~β_2 are regression coefficients. Average values of fitted parameters are given in the results section where applicable. For behavioral analysis, we fit these models separately to each session (182 for monkey U, 134 for monkey X). To examine the relationship between model predictions and values from regression analysis of neural data, we fit these models separately for the trials associated with each neuronal recording within each session. For these analyses, sample size corresponds to the number of neurons included in the regression analysis.

Linear Analysis of Neural Data—All analysis of neural data was conducted with custom software written in MATLAB. For all regression analyses, we analyzed activity during the fore-period (−1 to 0 s before target onset) for signals related to the previous trial, and the post-target period (0.25 s to 1.25 s after target onset) for signals pertinent to the stimuli of the current trial. The former epoch was chosen because it is the period immediately prior to the onset of targets during the current trial, and is thus the last window before signals related to the previous trial could be affected by information about the current trial. The latter epoch was chosen because it contains the majority of the time during the current trial, plus a small 0.25 s delay for stimulus information to reach the PFC. The spike density functions were constructed with a Gaussian filter (σ=40ms) and all other visualizations of the results of the regression were obtained with a 0.5 s sliding window. Neurons for which the design matrix was not full rank in either volatile or stable were omitted from the analysis (see below).

A multiple linear regression model was used to determine how individual neurons in each region encoded various types of information in the volatile and stable blocks. Some of this information was obtained by combining more basic elementary terms. For example, the color of the previously chosen target and the current positions of the two target colors can be combined to indicate the current location of the previously chosen color. To analyze how such higher-order features were encoded by single neurons, we compared the fit of two regression models to activity in the volatile and stable blocks. The following model was fit to the firing rate of each neuron during the volatile block.

$$y(t) = \beta_0 + \beta_1 \Delta M_{CU}(t) + \beta_2 \Delta M_{LR}(t) + \beta_3 C_{LR}(t-1) + \beta_4 C_{LR}(t) + \beta_5 C_{RG}(t-1) + \beta_6 C_{RG}(t) + \beta_7 R(t-1) + \beta_8 POS_{RG}(t) + \beta_9 HV_{RG}(t) + \beta_{10} C_{LR}(t-1) R(t-1) + \beta_{11} C_{RG}(t-1) R(t-1) + \beta_{12} C_{RG}(t-1) POS_{RG}(t) + \beta_{13} POS_{RG}(t) R(t-1) + \beta_{14} HV_{RG}(t) C_{RG}(t-1) + \beta_{15} HV_{RG}(t) C_{RG}(t) + \beta_{16} HV_{RG}(t) C_{RG}(t-1) R(t-1) + \beta_{17} HV_{RG}(t) C_{RG}(t-1) POS_{RG}(t) + \beta_{18} HV_{RG}(t) POS_{RG}(t) R(t-1) + \beta_{19} HV(t),$$

where $y(t)$ is the firing rate of a neuron for a given epoch in trial t, $\Delta M_{CU}(t)$ is the difference in reward magnitude between the chosen and unchosen targets in trial t, $\Delta M_{LR}(t)$ is the difference in reward magnitude between the left and right targets in trial t, $C_{LR}(t)$ is the position of the chosen target in trial t (1 if right, −1 otherwise), $C_{RG}(t)$ is the chosen target color in trial t (1 if red, −1 otherwise), $R(t)$ is the outcome in trial t (1 if rewarded, −1 otherwise), and $POS_{RG}(t)$ is the position of the red and green targets on trial t (1 if red is on the right, −1 otherwise). The $HV_{RG}(t)$ term indicates the target currently associated with the high reward probability target (1 if red is high probability, −1 otherwise). The last variable in the model, $HVL(t)$, is the three way interaction given by $C_{RG}(t{-}1)$ $R(t{-}1)$ $POS_{RG}(t)$. This indicates the current position of the target color that was rewarded in the previous trial. For example, if the animal chose red (green) on the previous trial and was rewarded (unrewarded), and red was on the left in the current trial, then the current position of the previously rewarded target would be on the left. $\beta_0 - \beta_{19}$ are regression coefficients.

During the stable blocks, target features were strictly correlated with reward probabilities. Thus, it was necessary to use similar but separate models for the volatile and stable blocks to account for these differences. The following model was fit to the firing rate of each neuron during the stable block.

$$y(t) = \beta_0 + \beta_1 \Delta M_{CU}(t) + \beta_2 \Delta M_{LR}(t) + \beta_3 C_{LR}(t-1) + \beta_4 C_{LR}(t) + \beta_5 C_{HV}(t-1) + \beta_6 C_{HV}(t) + \beta_7 R(t-1) + \beta_8 POS_{HV}(t) + \beta_9 B_{CS}(t) + \beta_{10} C_{LR}(t-1) R(t-1) + \beta_{11} C_{HV}(t-1) R(t-1) + \beta_{12} C_{HV}(t-1) POS_{HV}(t) + \beta_{13} POS_{HV}(t) R(t-1) + \beta_{14} B_{CS}(t) C_{HV}(t-1) + \beta_{15} B_{CS}(t) C_{HV}(t) + \beta_{16} B_{CS}(t) C_{HV}(t-1) R(t-1) + \beta_{17} B_{CS}(t) C_{HV}(t-1) POS_{HV}(t) + \beta_{18} B_{CS}(t) POS_{HV}(t) R(t-1) + \beta_{19} HVL(t),$$

where $C_{HV}(t)$ is the chosen target in trial t (1 if high-reward-probability target, −1 otherwise) and $POS_{HV}(t)$ is the position of the high- and low-reward-probability targets on trial t (1 if the high-reward-probability target is on the right, −1 otherwise). Additionally, the $B_{CS}(t)$ term indicates whether the targets are differently colored or shaped on trial t. Because target features are correlated with reward probability during the stable block, the $C_{HV}(t{-}1)$ term in this model is analogous to the $C_{RG}(t{-}1)HV_{RG}(t)$ term in the above model for volatile blocks, and the $B_{CS}(t)$ $C_{HV}(t)$ term indicates the position of a specific target feature, analogous to the $C_{RG}(t)$ term. Remaining terms have the same meaning as in the model for volatile blocks. For example, the $HVL(t)$ is the interaction between $CHV(t{-}1)$, $R(t{-}1)$, and $POS_{HV}(t)$, and therefore indicates the position of the target color (or shape) that yielded a reward on the previous trial. For the analysis of reward signals during the feedback period, we added a term indicating whether the outcome of the current trial was rewarded or not, $R(t)$, to the above two models.

Although we report the results only for a subset of the variables in this model, the inclusion of each term was necessary. Some terms corresponded to a lower-order effect associated with HVL, so they were necessary for interpreting values and significance of the corresponding regression coefficient. Others were included because they are known to have a neural representation in the prefrontal cortex (Donahue and Lee, 2015; Kim et al., 2008; Barraclough et al., 2004), and is thus an important factor to control for when analyzing the effects of other factors on firing rates. We omitted neurons from the analysis for which some regressors were perfectly correlated in either model, or there were an insufficient number of spikes to fit the model during the time windows of interest. For the fore-period, this left 164 (94.3%) neurons in the DLPFC, 132 (97.8%) neurons in the OFC, and 121 (89.6%) neurons in the ACC that could be fitted by both models. For the post-target period, 162 (93.1%) neurons in the DLPFC, 131 (97.0%) neurons in the OFC, and 119 (88.1%) neurons in the ACC could be fitted by both models.

After fitting, the coefficient of partial determination (CPD) for each variable was computed separately for individual neurons, as follows.

$$CPD = \frac{SSE_{reduced} - SSE_{full}}{SSE_{reduced}},$$

where SSE_{full} is the sum of squared errors for the full model, and $SSE_{reduced}$ is the sum of squared errors for a reduced model that omits the variable of interest. CPD quantifies how much variance in the data is accounted for by that variable. We performed a mixed-effects ANOVA on the CPD to test whether there were volatility effects or regional differences in the representation of each variable, by treating volatility as a within-subject factor and region as a between-subject factor. Raw CPDs are shown in figures, but they were log-transformed prior to conducting statistical analysis. The mean and S.E.M. of all CPDs tested is given in figures.

To analyze the representations of variables of interest within individual neurons, we performed 2-way ANOVAs with the variable of interest, and volatility as factors. One exception was the analysis of the fraction of neurons that significantly encoded the difference between chosen and unchosen reward magnitudes, for which we used a simple multiple regression model with volatility and the difference in chosen magnitude as factors. ANOVA was not possible for this analysis, because the animals seldom chose the low-reward-probability target with small reward magnitude during the stable blocks. Mean and S.E.M. of firing rate are depicted in figures, and the number of trials is given in the main text.

Single Neuron Decoding Analysis—We applied a linear discriminant analysis with an *n*-fold cross-validation to determine whether neural signals related to spatial position, volatility, and target color were modulated by previous outcomes (Donahue and Lee, 2015). The classifiers were trained to classify a particular variable (discriminandum) using the firing rates of an individual neuron during a given time window. We trained classifiers to decode the previously chosen location, the location of the upcoming choice, the previously chosen target color (or shape), the position of the two targets, and the current location of the previously chosen target color (or shape). Trials were randomly assigned to *n* different subgroups for the *n*-fold cross-validation. Each subgroup served as a test set once, with the trials in the remaining subgroups used as the training set. To ensure that classifiers were unbiased, each subgroup was balanced by randomly removing trials until each value of the discriminanda was equally frequent. Additionally, the subgroups for the classifier decoding the location of the upcoming choice were balanced with respect to the current position of the previously rewarded target color, and the subgroups for the classifier decoding the current location of the previously chosen color and the classifier decoding the position of the two targets were also balanced with respect to the previously chosen target color (or shape).

To obtain a sufficient number of neurons available for the analysis, a 5-fold cross-validation was used for decoding the previously chosen location and location of the upcoming choice, and a 3-fold cross validation was used for decoding the previously chosen target color, the position of the two targets, and the current location of the previously chosen target color. Neurons that had fewer than *n* trials (needed for an *n*-fold cross-validation) for any combination of the discriminandum and other variables used for balancing were thus removed from the analysis. For example, the classifier trained to decode the current position of the previously chosen target color required at least 3 trials for each of the 16 possible combinations of the position of the previously chosen target, the previously chosen target color (or shape), and the previous outcome, and block type (volatile vs. stable). This resulted in 88 DLPFC neurons, 90 OFC neurons, and 84 ACC neurons in the decoding analysis for the current location of the previously chosen target and the position of the two targets. Similarly, 166 DLPFC neurons, 132 OFC neurons, and 129 ACC neurons were included in the decoding analysis for the location of the upcoming choice, whereas 168 DLPFC neurons, 132 OFC neurons, and 128 ACC neurons were included in the decoding analysis for the previously chosen target color. All recorded neurons were included in the decoding analysis for the previously chosen location. A mixed three-way ANOVA with region, reward, and volatility as factors was used to examine the representation of these variables in the prefrontal cortex during the post-target period (0.25 s to 1.25 s after target presentation). Further, a repeated measures two-way ANOVA was performed on decoding accuracy estimated for the post-target period for each classifier with reward and volatility as factors.

The mean and S.E.M. of decoding accuracies used for statistical analysis is given by bar plots in corresponding figures. For visualization, decoding accuracy was also performed in a 1-s sliding-window advancing in 50-ms steps. The subgroup sampling and cross-validation procedure was repeated 100 times for each classifier, and the decoding accuracy was averaged across these repetitions. We also directly tested whether signals related to the current position of each target color in the stable block and the current position of the previously rewarded target color (HVL) in the volatile block were affected differently by the outcome of the previous trial. To do this, we compared the decoding accuracy for the position of the high-reward-probability target in stable block following rewarded and unrewarded trials to the decoding accuracy for the current position of the previously rewarded target color in volatile block, using a repeated measures two-way ANOVA with reward and volatility as factors.

DATA AND SOFTWARE AVAILABILITY

Custom software used in the above analyses will be made available upon reasonable request to the Lead Contact.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank M. Hammond, M. Gay, J. Schnurr, V. Zafrany, I. Bobeica, and M. Petrowicz for their technical assistance. This study was supported by the National Institute of Health (T32 EY022312 to B. M; T32 NS007224 to C.H.D.; P30 EY026878, R01 DA029330 and MH 108629 to D.L.).

References

Asaad WF, Lauro PM, Perge JA, Eskandar EN. Prefrontal Neurons Encode a Solution to the Credit-Assignment Problem. J. Neurosci. 2017; 37:6995–7007. [PubMed: 28634307]

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 TE, Woolrich MW, Walton ME, Rushworth MF. 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]

Cai X, Padoa-Schioppa C. Contributions of orbitofrontal and lateral prefrontal cortices to economic choice and the good-to-action transformation. Neuron. 2014; 81:1140–1151. [PubMed: 24529981]

Camille N, Tsuchida A, Fellows LK. Double dissociation of stimulus-value and action-value learning in humans with orbitofrontal or anterior cingulate cortex damage. J. Neurosci. 2011; 31:15048–15052. [PubMed: 22016538]

Chau BK, Sallet J, Papageorgiou GK, Noonan MP, Bell AH, Walton ME, Rushworth MF. Contrasting roles for orbitofrontal cortex and amygdala in credit assignment and learning in macaques. Neuron. 2015; 87:1106–1118. [PubMed: 26335649]

Chaudhuri R, Knoblauch K, Gariel MA, Kennedy H, Wang XJ. A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. Neuron. 2015; 88:419–431. [PubMed: 26439530]

Compte A, Brunel N, Goldman-Rakic PS, Wang XJ. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. Cereb. Cortex. 2000; 10:910–923. [PubMed: 10982751]

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

Donahue CH, Lee D. Dynamic routing of task-relevant signals for decision making in dorsolateral prefrontal cortex. Nat. Neurosci. 2015; 18:295–301. [PubMed: 25581364]

Donahue CH, Seo H, Lee D. Cortical signals for rewarded actions and strategic exploration. Neuron. 2013; 80:223–234. [PubMed: 24012280]

Doya K. Metalearning and neuromodulation. Neural Netw. 2002; 15:495–506. [PubMed: 12371507]

Fellows LK, Farah MJ. Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm. Brain. 2003; 126:1830–1837. [PubMed: 12821528]

Farashahi S, Donahue CH, Khorsand P, Seo H, Lee D, Soltani A. Metaplasticity as a neural substrate for adaptive learning and choice under uncertainty. Neuron. 2017; 94:401–414. [PubMed: 28426971]

Fusi S, Miller EK, Rigotti M. Why neurons mix: high dimensionality for higher cognition. Curr. Opin. Neurobiol. 2016; 37:66–74. [PubMed: 26851755]

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

Hayden BY, Platt ML. Neurons in anterior cingulate cortex multiplex information about reward and action. J. Neurosci. 2010; 30:3339–3346. [PubMed: 20203193]

Heilbronner SR, Hayden BY. Dorsal anterior cingulate cortex: a bottom-up view. Annu. Rev. Neurosci. 2016; 39:149–170. [PubMed: 27090954]

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]

Ito M, Doya K. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. J. Neurosci. 2009; 29:9861–9874. [PubMed: 19657038]

Jones JL, Esber GR, McDannald MA, Gruber AJ, Hernandez A, Mirenzi A, Schoenbaum G. Orbitofrontal cortex supports behavior and learning using inferred but not cached values. Science. 2012; 338:953–956. [PubMed: 23162000]

Kazama A, Bachevalier J. Selective aspiration or neurotoxic lesions of orbital frontal areas 11 and 13 spared monkeys' performance on the object discrimination reversal task. J. Neurosci. 2009; 29:2794–2804. [PubMed: 19261875]

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

Kim S, Cai X, Hwang J, Lee D. Prefrontal and striatal activity related to values of objects and locations. Front. Neurosci. 2012; 6:108. [PubMed: 22822390]

Kim S, Hwang J, Lee D. Prefrontal coding of temporally discounted values during intertemporal choice. Neuron. 2008; 59:161–172. [PubMed: 18614037]

Lee SW, Shimojo S, O’Doherty JP. Neural computations underlying arbitration between model-based and model-free learning. Neuron. 2014; 81:687–699. [PubMed: 24507199]

McGuire JT, Nassar MR, Gold JI, Kable JW. Functionally dissociable influences on learning rate in a dynamic environment. Neuron. 2014; 84:870–881. [PubMed: 25459409]

Meder D, Kolling N, Verhagen L, Wittmann MK, Scholl J, Madsen KH, Hulme OJ, Behrens TEJ, Rushworth MF. Simultaneous representation of a spectrum of dynamically changing value estimates during decision making. Nat. Commun. 2017; 8:1942. [PubMed: 29208968]

Miller EK, Cohen JD. An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci. 2001; 24:167–202. [PubMed: 11283309]

Murray JD, Anticevic A, Gancsos M, Ichinose M, Corlett PR, Krystal JH, Wang XJ. Linking microcircuit dysfunction to cognitive impairment: effects of disinhibition associated with schizophrenia in a cortical working memory model. Cereb. Cortex. 2012; 24:859–872. [PubMed: 23203979]

Murray JD, Bernacchia A, Freedman DJ, Romo R, Wallis JD, Cai X, Padoa-Schioppa C, Pasternak T, Seo H, Lee D, Wang XJ. A hierarchy of intrinsic timescales across primate cortex. Nat. Neurosci. 2014; 17:1661–1663. [PubMed: 25383900]

Nassar MR, Rumsey KM, Wilson RC, Parikh K, Heasly B, Gold JI. Rational regulation of learning dynamics by pupil-linked arousal systems. Nat. Neurosci. 2012; 15:1040–1046. [PubMed: 22660479]

Padoa-Schioppa C, Assad JA. Neurons in orbitofrontal cortex encode economic value. Nature. 2006; 441:223. [PubMed: 16633341]

Rudebeck PH, Behrens TE, Kennerley SW, Baxter MG, Buckley MJ, Walton ME, Rushworth MF. Frontal cortex subregions play distinct roles in choices between actions and stimuli. J. Neurosci. 2008; 28:13775–13785. [PubMed: 19091968]

Rudebeck PH, Saunders RC, Prescott AT, Chau LS, Murray EA. Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. Nat. Neurosci. 2013; 16:1140–1145. [PubMed: 23792944]

Schoenbaum G, Nugent SL, Saddoris MP, Setlow B. Orbitofrontal lesions in rats impair reversal but not acquisition of go, no-go odor discriminations. Neuroreport. 2002; 13:885–890. [PubMed: 11997707]

Seamans JK, Durstewitz D, Christie BR, Stevens CF, Sejnowski TJ. Dopamine D1/D5 receptor modulation of excitatory synaptic inputs to layer V prefrontal cortex neurons. Proc. Natl. Acad. Sci. U.S.A. 2001; 98:301–306. [PubMed: 11134516]

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]

Soltani A, Noudoost B, Moore T. Dissociable dopaminergic control of saccadic target selection and its implications for reward modulation. Proc. Natl. Acad. Sci. U.S.A. 2013; 110:3579–3584. [PubMed: 23401524]

Sutton RS, Barto AG. Reinforcement Learning: An Introduction (MIT Press). 1998

Tervo DG, Proskurin M, Manakov M, Kabra M, Vollmer A, Branson K, Karpova AY. Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. Cell. 2014; 159:21–32. [PubMed: 25259917]

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

Walton ME, Behrens TE, Buckley MJ, Rudebeck PH, Rushworth MF. Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. Neuron. 2010; 65:927–939. [PubMed: 20346766]

Wang XJ. Decision making in recurrent neuronal circuits. Neuron. 2008; 60:215–234. [PubMed: 18957215]

Wang M, Yang Y, Wang CJ, Gamo NJ, Jin LE, Mazer JA, Morrison JH, Wang XJ, Arnsten AF. NMDA receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex. Neuron. 2013; 77:736–749. [PubMed: 23439125]

Wong KF, Wang XJ. A recurrent network mechanism of time integration in perceptual decisions. J. Neurosci. 2006; 26:1314–1328. [PubMed: 16436619]

Young CE, Yang CR. Dopamine D1-like receptor modulates layer-and frequency-specific short-term synaptic plasticity in rat prefrontal cortical neurons. Eur. J. Neurosci. 2005; 21:3310–3320. [PubMed: 16026469]

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

HIGHLIGHTS

- Animals repeat rewarded actions more often when reward probabilities vary.
- Outcome signals are stronger in the OFC when reward probabilities vary.
- Rewards enhance task-relevant signals in the DLPFC when reward probabilities vary.
- Task-irrelevant signals are unaffected by the outcome of the previous trial.

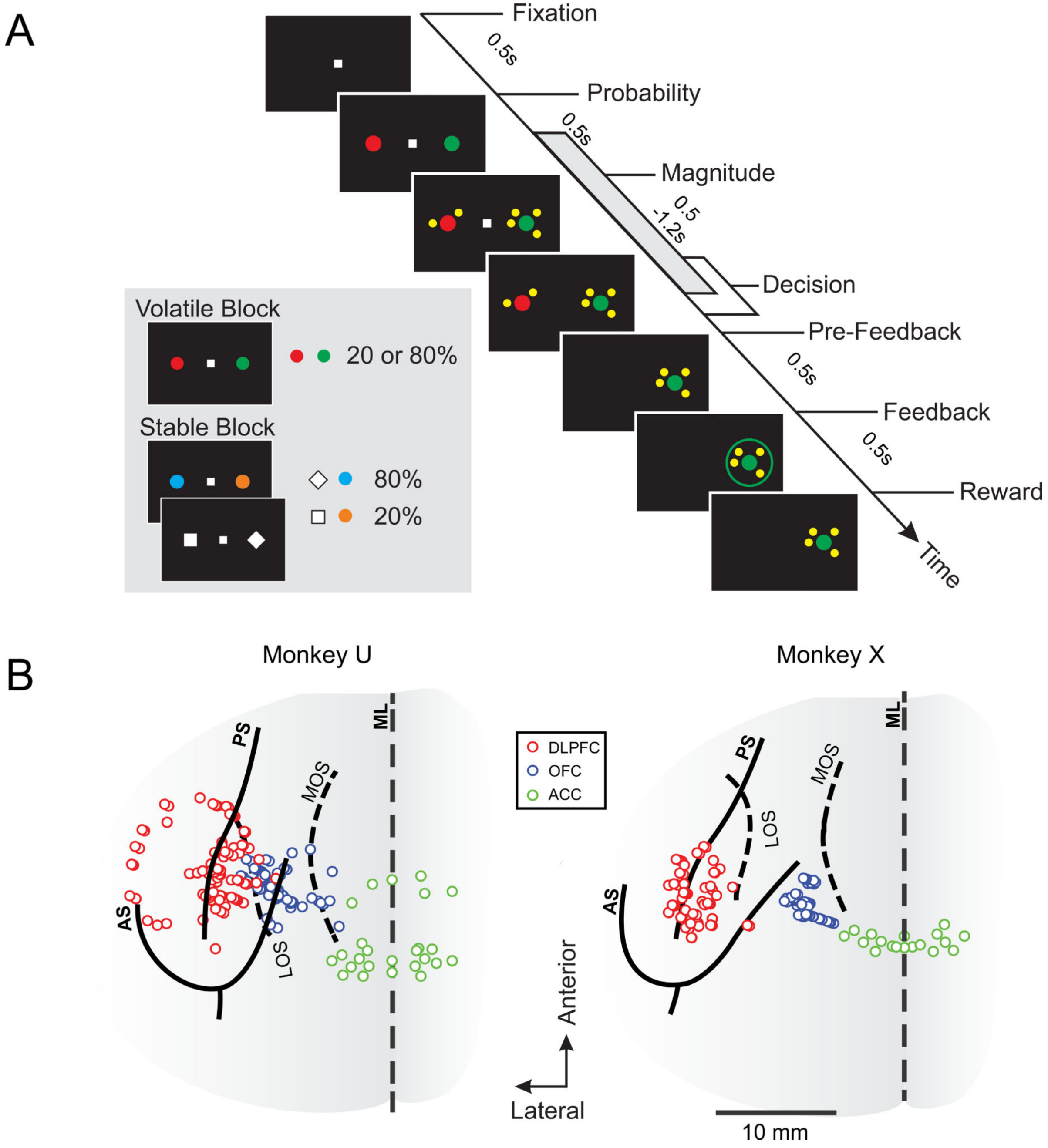


Figure 1. Probabilistic reversal-learning task

(A) The interval used for decoding analysis is illustrated as the grey bar along the time axis. Targets used for the volatile and stable trials and the corresponding reward probabilities are shown in the inset.

(B) The solid and dashed lines represent the sulci on the dorsal (AS, arcuate sulcus; PS, principal sulcus) and ventral (LOS, lateral orbital sulcus; MOS, medial orbital sulcus) surfaces of the brain, respectively. ML, midline. All recording sites in the ACC were located dorsal to the cingulate sulcus in both animals.

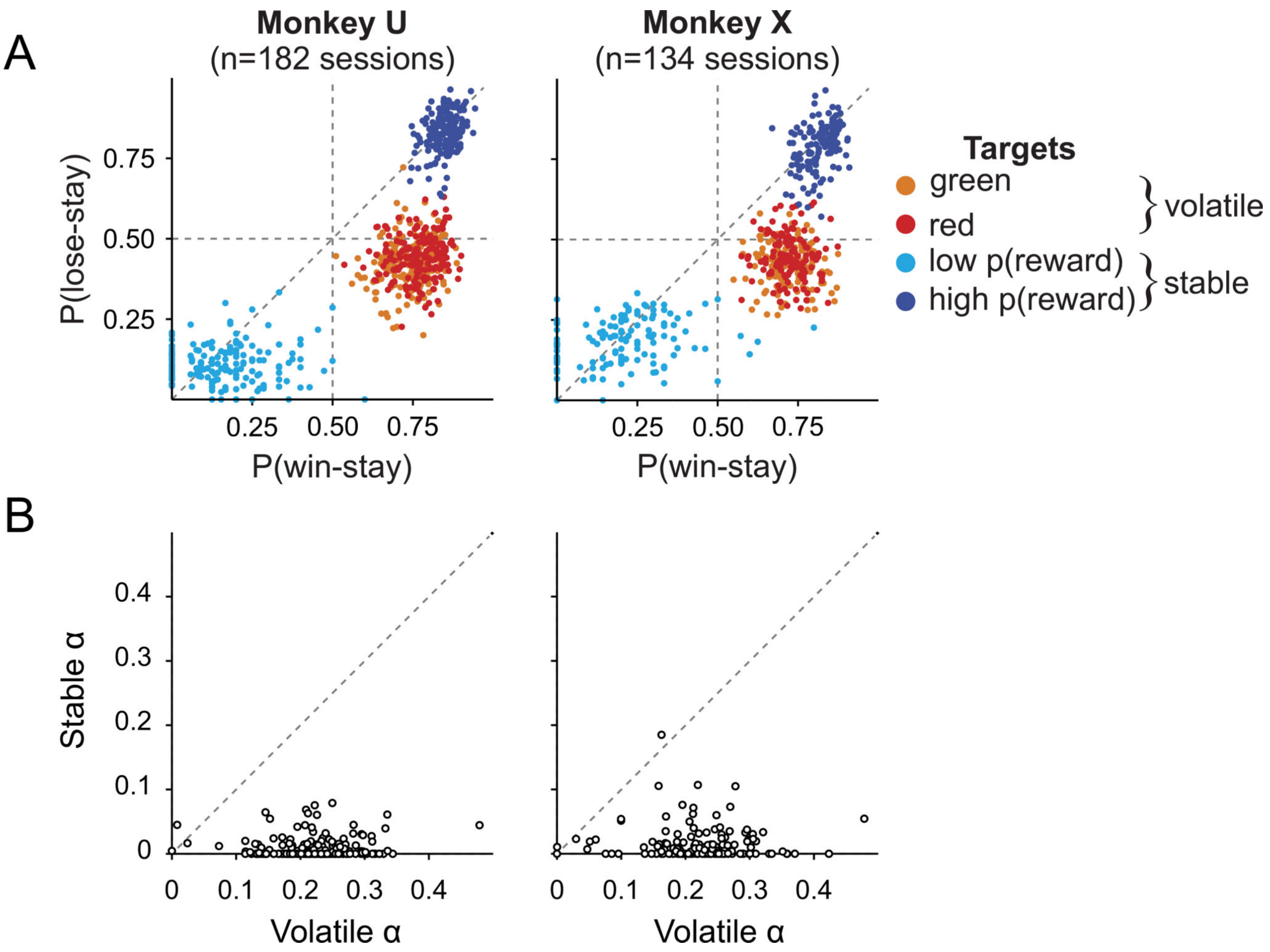


Figure 2. Behavioral Performance

(A) The proportion of trials in which each animal chose the same target color (or shape) as in the previous trial after the previous choice was rewarded (win-stay) or unrewarded (lose-stay), separately for different targets in the volatile and stable blocks.

(B) Learning rates (α) in the volatile and stable block for each animal (see STAR Methods). See also Figure S1.

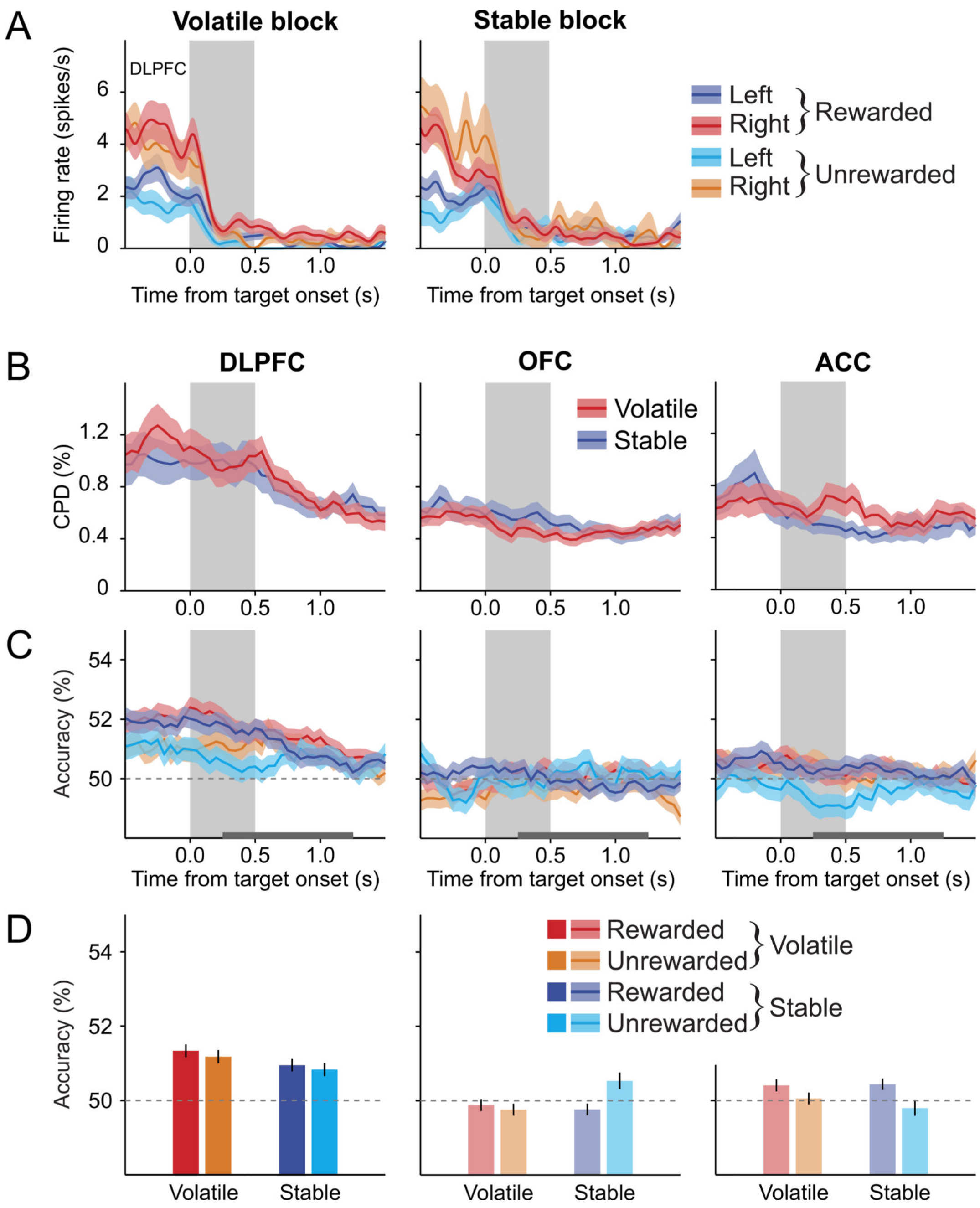


Figure 3. Signals related to previous action in the PFC are unaffected by previous outcome

(A) An example DLPFC neuron that encoded the action in the previous trial during the fore-period (n = 591 trials, effect of previously chosen location in a 3-way ANOVA, $p<10^{-7}$). The effect of volatility or previous reward was not significant (p>0.2).

(B) Time course of the mean CPD for the previously chosen target location. Gray background indicates the target period.

(C) Time course of the average decoding accuracy for previously chosen target location, plotted separately according to block types and previous outcomes.

(D) Average decoding accuracy for the previously chosen target location during the post-target period (horizontal gray line in B). Shaded areas in (A) and (B) and error bars in (C) represent ± SEM. Lighter bars in (C) indicate that the decoding accuracies were not significantly above the chance level (one-sample t-test, p>0.05). See also Figure S2.

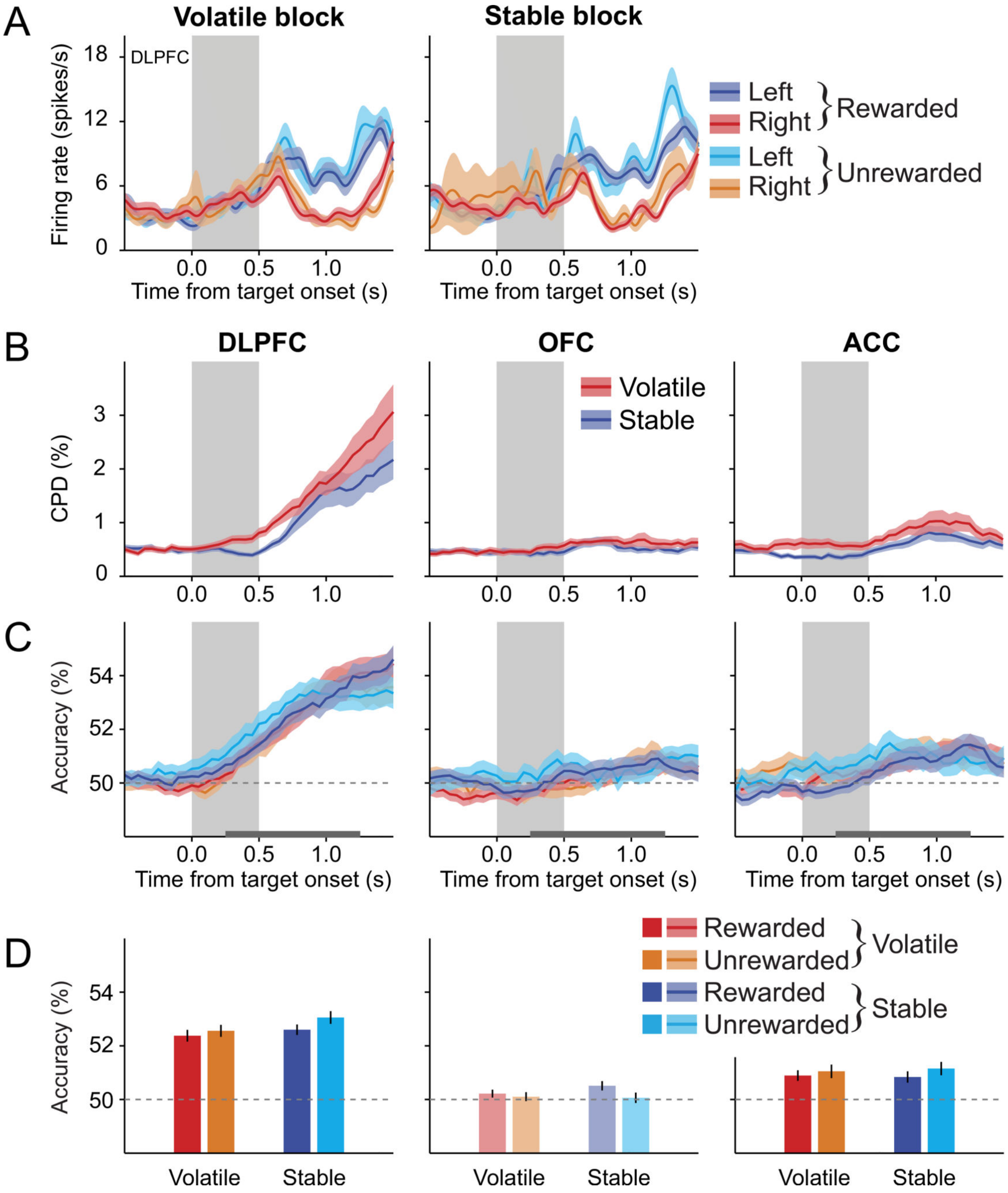


Figure 4. Signals related upcoming action in the PFC is unaffected by previous outcome

(A) An example DLPFC neuron that encoded the animal’s current choice during the 0.5-s interval beginning 0.25 s after target onset (n = 640 trials, effect of chosen target location in a 3-way ANOVA, p<0.01), but without significant effect of volatility (p>0.1) or previous reward (p>0.05).

(B) Time course of the mean CPD for the position of the target chosen in the current trial.

(C) Time course of the average decoding accuracy for the position of the target chosen in the current trial, plotted separately according to block types and previous outcomes.

(D) Average decoding accuracy for the position of the target chosen in the current trial during the post-target period (horizontal gray line in (B)). Same format as in Figure 3. See also Figure S3.

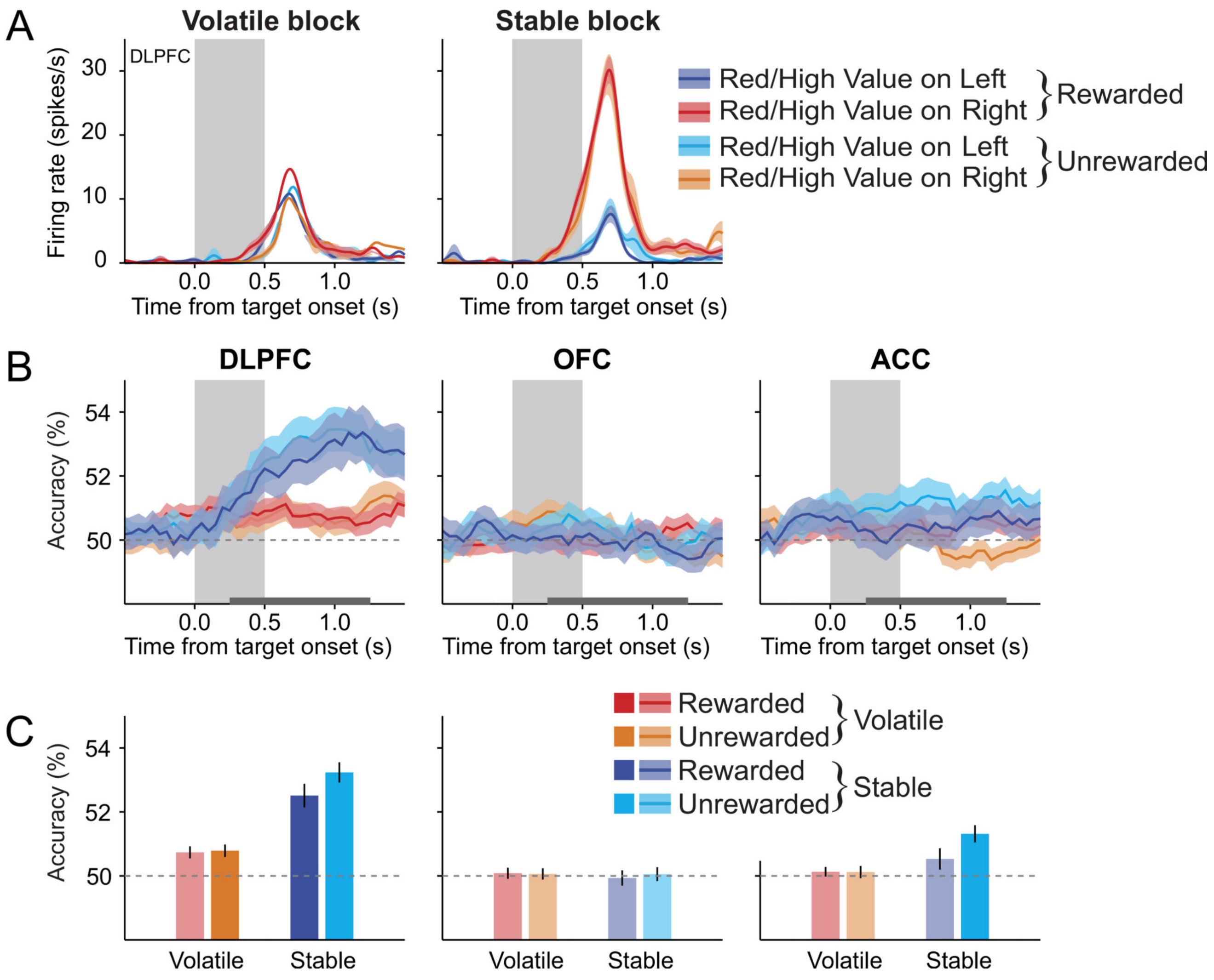


Figure 5. Signals related to the position of the high-reward probability target in stable blocks are not affected by previous outcome

(A) An example DLPFC neuron that significantly changed its activity according to the position of the high reward-probability target in stable blocks (n = 330 trials, effect of target position in a 3-way ANOVA, $p<0.10^{-18}$) significantly more strongly during the stable blocks (volatility × target position interaction, $p<10^{-15}$). The activity of this neuron was not affected by previous reward (main effect of reward and its interactions, p>0.05).

(B) Time course of the average decoding accuracy for the relative positions of the two different target colors (or shapes) following rewarded and unrewarded trials in the volatile and stable blocks. Shaded areas represent ± SEM.

(C) Average decoding accuracy for the relative positions of the two different target colors (or shapes) during the post-target period (horizontal gray line in (A)). Error bars represent ± SEM. Lighter bars indicate that the decoding accuracies were not significantly above the chance level (one-sample t-test, p>0.05). See also Figure S4.

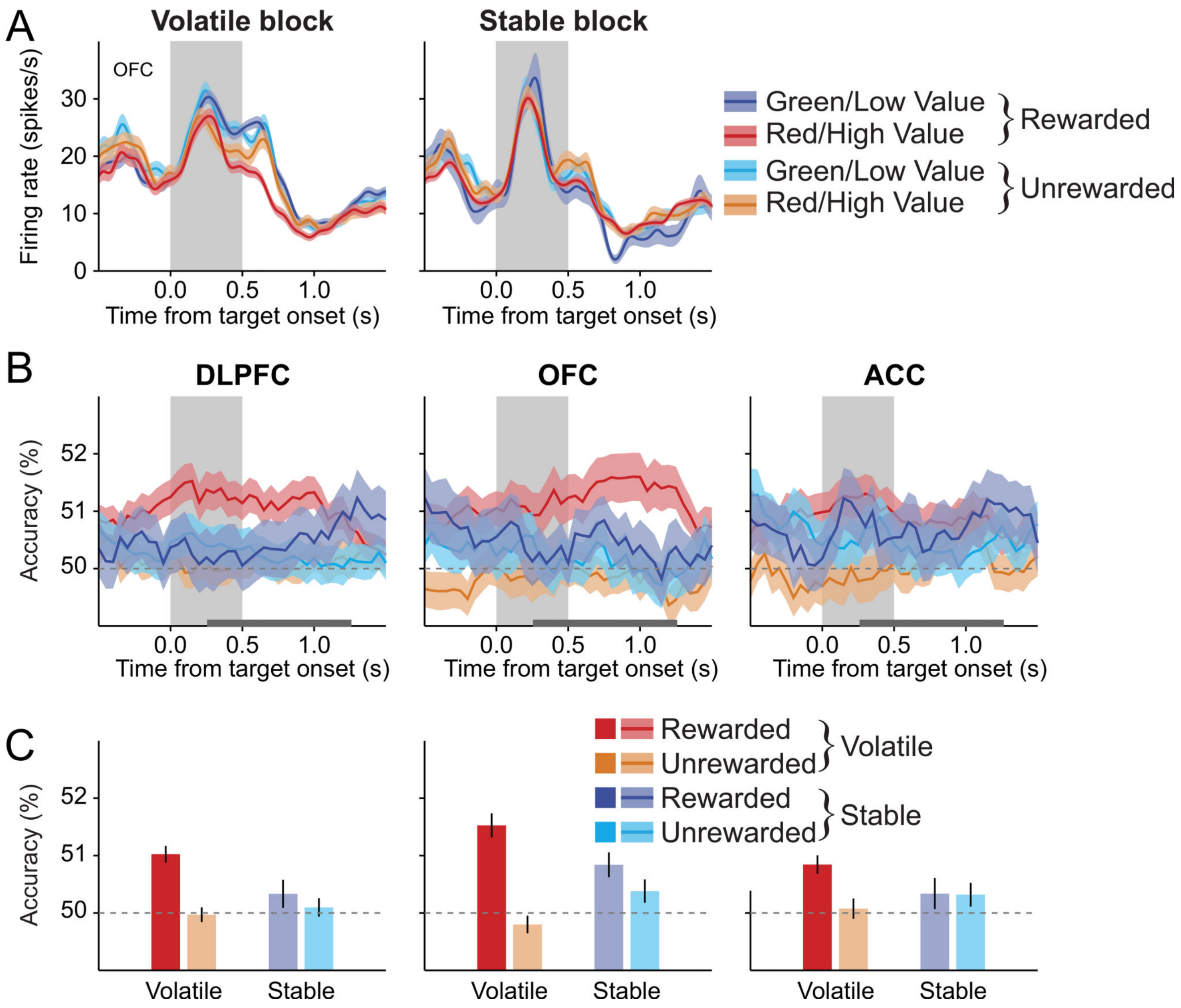


Figure 6. Signals related to the target color or shape chosen in the previous trial in the PFC

(A) An example DLPFC neuron that significantly encoded the previously chosen target color (or shape) during the 0.5-s interval beginning 0.25 s after target onset, but only when this was rewarded in volatile blocks (n = 640 trials, previously chosen target × reward and previously chosen target × volatility interactions in a 3-way ANOVA, p<0.05).

(B)Time course of the average decoding accuracy for the previously chosen color (or shape), plotted separately according to block types and previous outcomes.

(C) Average decoding accuracy for the previously chosen color during the post-target period (horizontal gray line in (A)). Same format as in Figure 5. See also Figure S5.

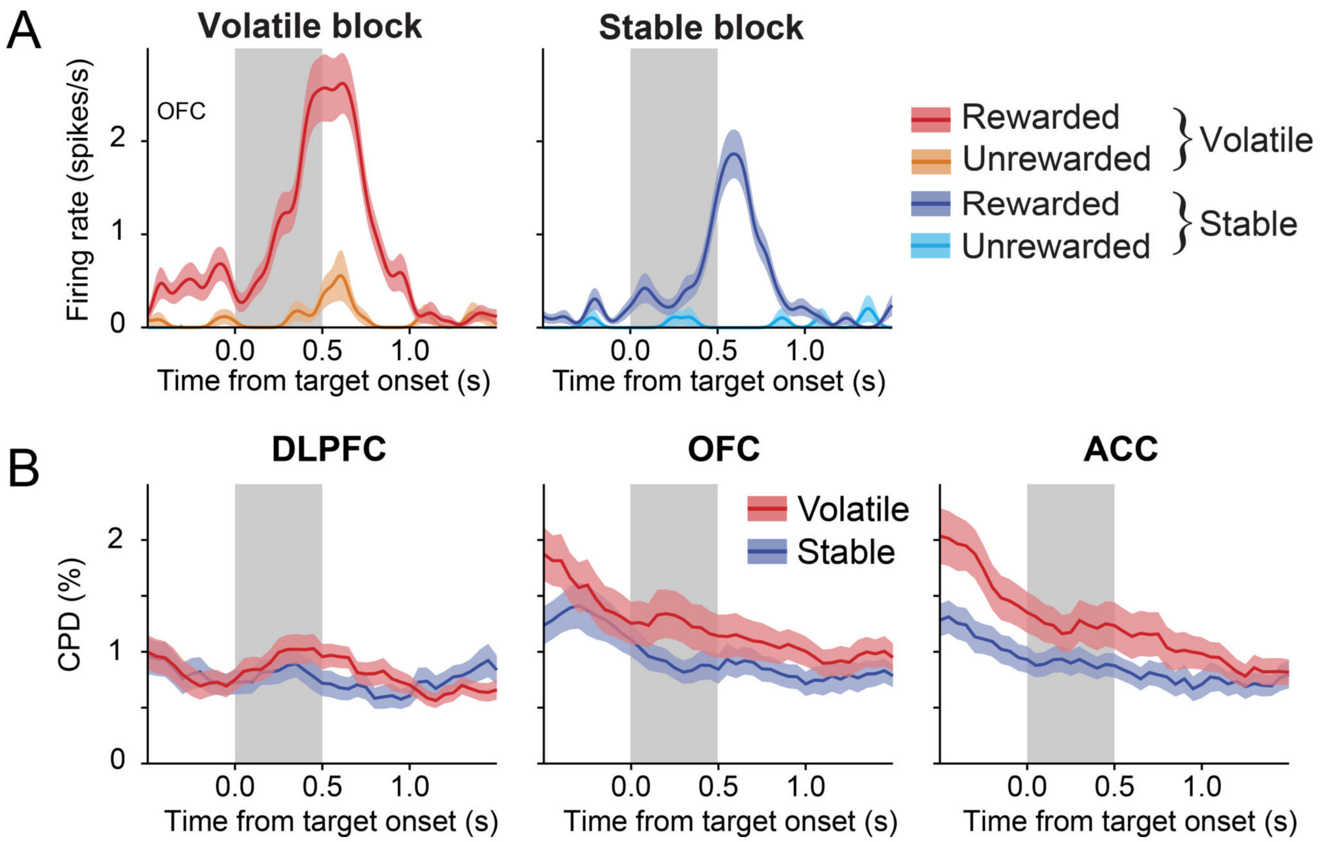


Figure 7. Effect of volatility on reward signals in the PFC

(A) An example OFC neuron that encoded the outcome of the previous trial more strongly during the fore-period in the volatile block than in the stable block.

(B) Time course of the CPD for the reward in the previous trial, shown separately for stable and volatile blocks. Shaded areas represent ± SEM. See also Figure S6.

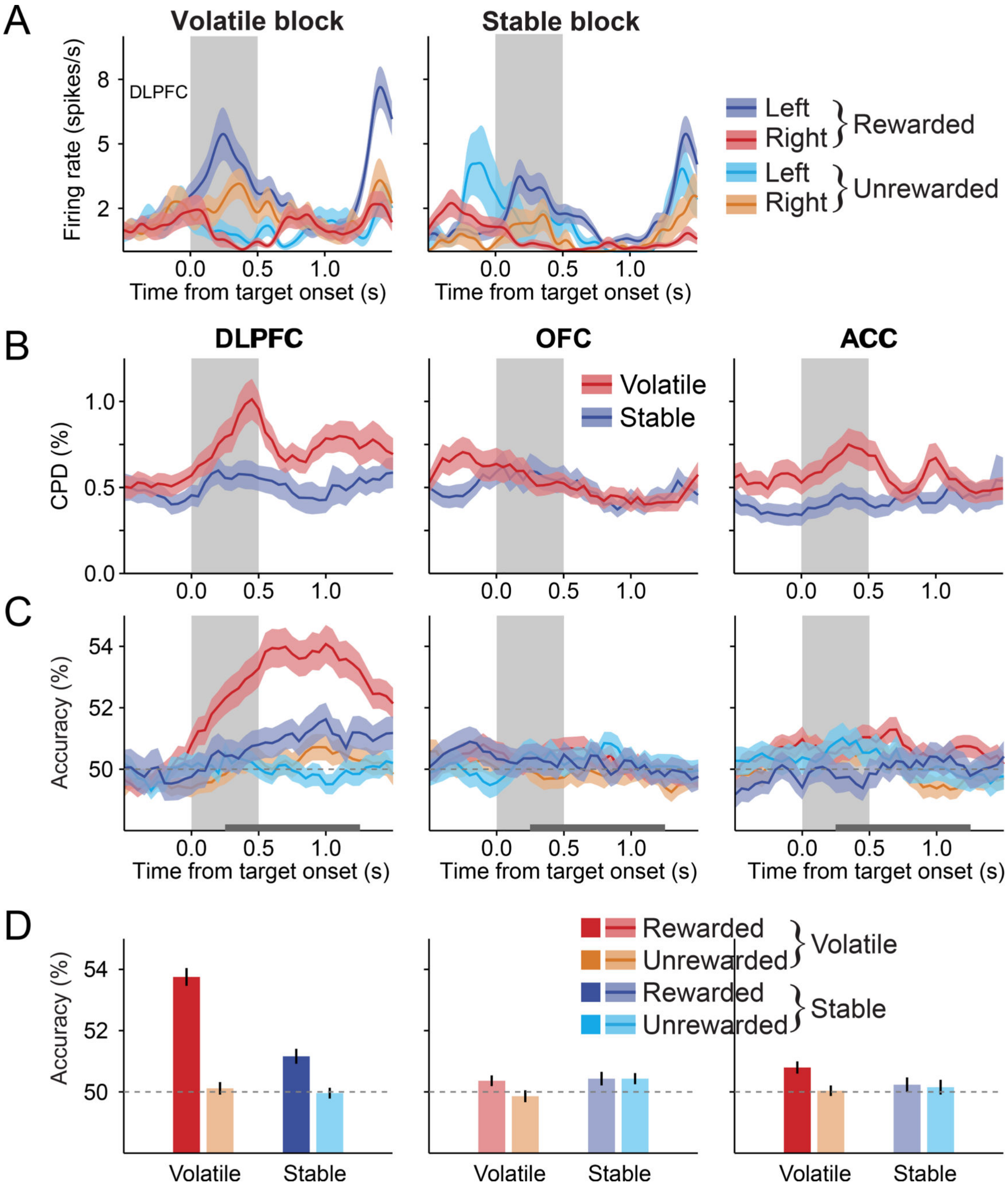


Figure 8. Effect of previous reward and volatility on signals related to the current position of the previously chosen target color or shape

(A) An example DLPFC neuron that encoded the current location of the previously chosen target color (or shape) during the target period (gray background) more strongly when the previous choice was rewarded compared to when it was not in the volatile block, but not the stable block.

(B) Time course of the mean CPD for the current position of the target color or shape that was or would have been rewarded in the previous trial (HVL).

(C) Time course of the average decoding accuracy for the current position of the previously chosen target color (or shape), plotted separately according to block types and previous outcomes.

(D) Average decoding accuracy for the current position of the previously chosen target color (or shape) during the post-target period (horizontal gray line in (C)). Shaded areas in (A)-(C) and error bars in (D) represent ± SEM. Lighter bars indicate that the decoding accuracies were not significantly above the chance level (one-sample t-test, $p>0.05$). See also Figure S7.