Published in final edited form as:
Science. 2014 October 17; 346(6207): 340–343. doi:10.1126/science.1256254.

Neural Correlates of Strategic Reasoning during Competitive Games◆

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

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

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

[3]Department of Psychology, Yale University, New Haven, CT06510, USA

Abstract

Although human and animal behaviors are largely shaped by reinforcement and punishment, choices in social settings are also influenced by information about the knowledge and experience of other decision-makers. During competitive games, monkeys increased their payoffs by systematically deviating from a simple heuristic learning algorithm and thereby countering the predictable exploitation by their computer opponent. Neurons in the dorsomedial prefrontal cortex (dmPFC) signaled the animal's recent choice and reward history that reflected the computer's exploitative strategy. The strength of switching signals in the dmPFC also correlated with the animal's tendency to deviate from the heuristic learning algorithm. Therefore, the dmPFC might provide control signals for overriding simple heuristic learning algorithms based on the inferred strategies of the opponent.

Learning algorithms suffer from the curse of dimensionality, as the amount of data necessary for statistically robust learning increases with the complexity of the task (1). Simple heuristic learning algorithms can thus be more effective, even for complex tasks (2). A broad range of animal and human behaviors follows model-free reinforcement learning algorithms operating with a small number of discrete states (3–5). Nevertheless, more complex learning models can be advantageous when sufficient knowledge of the decision maker's environment is available. In particular, inferences about the likely behaviors of other decision makers often complement simple learning algorithms in social settings (6–11). However, the nature of control signals resulting from complex strategic inferences and how they are incorporated into the process of action selection remain unknown.

*Corresponding authors (hyojung.seo@yale.edu, daeyeol.lee@yale.edu).
†Current address: NYU-ECNU Joint Institute of Brain and Cognitive Sciences, NYU-Shanghai, Shanghai, China.
‡Current address: Gladstone Institute of Neurological Disease, San Francisco, CA 94158., USA.

Supplementary Materials
www.sciencemag.org
Materials and Methods
Figs. S1 to S8.

To investigate the nature of neural signals responsible for disengaging the animals from simple heuristic learning, we trained three rhesus monkeys to perform a token-based oculomotor decision-making task modeled after a biased matching pennies game against a computer opponent (Fig. 1A) (12). The outcome in each trial was jointly determined by the choices of the animal and the computer opponent according to the payoff matrix of the game (Fig. 1B), and the number of tokens shown on the computer screen was adjusted after each trial accordingly. The animal gained a token whenever it chose the same target as the computer. When the animal and computer chose different targets, the outcome was neutral for one target (referred to as safe) and loss for the other target (referred to as risky). The animal received juice reward whenever it accumulated 6 tokens. The optimal strategy (known as Nash equilibrium (13)) for the animal was to choose the risky target with a probability of 1/3. If the animal chose the risky or safe target more frequently than predicted by the optimal strategy, this was exploited by the computer (14). The locations of risky and safe targets were fixed in a block of trials (mean = 47.6 ± 7.6 trials) and reversed between blocks. The computer used only the history of the animal's choices and outcomes in the current block to exploit the animal's biases.

The animals were more likely to choose the same target after receiving a token and to switch to the other target after losing a token, compared to when the outcome of their previous choice was neutral. The effects of gains and losses on subsequent choices decayed gradually (Fig. 1C). This is consistent with a model-free reinforcement learning algorithm in which the value functions for the two options are continually adjusted according to the outcome of the animal's choice (3–5). Such a simple learning algorithm during a competitive game might be disadvantageous because it leads to predictable choices. However, animals performed significantly better than the best-fitting model-free reinforcement learning algorithm and achieved payoffs indistinguishable from that of the Nash-equilibrium player (Fig. S1). This suggests that the animals might have complemented a model-free reinforcement learning algorithm with more flexible strategies, potentially counter-exploiting the computer's algorithm.

The animal's choices systematically deviated from the predictions from a model-free reinforcement learning algorithm. For some animals (monkeys H and J), this was manifest as the attenuation in the immediate effect of losing a token on the animal's behaviors (Fig. 1C, arrows). More generally, the animal's choices and their outcomes in the previous two trials reliably predicted whether the animal would choose the safe or risky target more frequently than predicted by a model-free reinforcement learning algorithm (Fig. 2), and this relationship was consistent across sessions within each animal, as well as across animals (Fig. S2). The computer opponent largely exploited the sequences of animal's choices and outcomes expected from simple reinforcement learning (Fig. S3A). The animals tended to choose the safe target much more frequently than predicted by the reinforcement learning algorithm, following the same sequence of outcomes that strongly biased the computer to predict the opposite (Fig. S3B). The computer's prediction for the animal's next choice was frequently exploited by the animals when it was based on simple patterns, such as the win-stay-lose-switch strategy (Fig. S3C). This increased the animal's expected payoff beyond the level predicted for the Nash-equilibrium strategy (Fig. 2).

We analyzed single-neuron activity recorded from the dorsolateral prefrontal cortex (dlPFC), dorsomedial prefrontal cortex (dmPFC), and dorsal anterior cingulate cortex (ACCd), as well as the dorsal (caudate nucleus, CD) and ventral striatum (VS; Fig. S4). Signals related to the animal's previous choices and their outcomes have been found in all of these areas, and can contribute to reinforcement learning (15–23). Because the animals tended to deviate from the model-free reinforcement learning frequently after specific sequences of choices and outcomes, we assumed that neurons involved in over-riding the use of simple reinforcement learning algorithm might encode high-order conjunctive signals related to the animal's choices and their outcomes in multiple trials. Neurons encoding such high-order conjunctions were common in the dmPFC. During the 0.5-s delay period, 48.7% of the neurons in the dmPFC significantly modulated their activity according to such conjunctions (14), and this was significantly higher than in all other areas tested in this study (χ^2-test, $p < 0.001$; Fig. 3).

We hypothesized that dmPFC activity related to conjunctions of previous choices and outcomes might provide the information necessary for the animal's decision to deviate from a model-free reinforcement learning algorithm. We decoded the animal's choice in the current trial from the activity of each neuron during the delay period, separately for different sequences of choices and outcomes in the 2 preceding trials, and quantified how much the accuracy of this decoding changed depending on whether the animal made the same choice in the previous trial or not (i.e., stay vs. switch; Fig. 4A) (14). We then tested whether this measure of switching activity was correlated with the frequency of deviations from model-free reinforcement learning across different outcome sequences (Fig. 4B). Significant correlation was found only for the dmPFC ($r = 0.14$, $p < 0.001$; Fig. 4C and 4D; Fig. S5 and S6). Therefore, switching activity in the dmPFC might contribute to strategically deviating from a simple reinforcement learning algorithm.

As a further independent test of this hypothesis, we applied the same methods to analyze the activity of neurons previously recorded from 4 cortical areas (dmPFC (23), dlPFC (21), ACC (16), and the lateral intaparietal area or LIP (22)) during an unbiased matching pennies task. Signals related to conjunctions of previous choices and outcomes were observed more frequently in the dmPFC than in other areas (χ^2-test, $p < 10^{-4}$; Fig. S7). Switching activity was also significantly correlated across different outcome sequences with the deviations from model-free reinforcement learning only for the dmPFC ($r = 0.13$, $p < 10^{-6}$; Fig. 4D; Fig. S8). These results suggest that activity in the dmPFC might be involved in multiple levels of switching, not only for switching between different motor responses (24–26), but also for switching between actions favored by simple reinforcement learning and more abstract strategic inferences.

For learning agents, the complexity of the optimal internal model for action selection not only depends on the complexity of the environment, but also increases with the amount of experience. Accordingly, animals learning in real time must apply multiple learning algorithms in parallel (4, 27–29). Real or simulated social interactions provide an ideal platform to investigate the dynamics and neural mechanisms for arbitrating between multiple learning algorithms. We found that monkeys can complement the use of a simple reinforcement learning algorithm with strategic high-order reasoning to improve their choice

outcomes during virtual competitive interaction. We also identified a neural signature of switching between different learning algorithms in the medial frontal cortex. Inabilities to choose appropriate learning algorithms are thought to underlie a number of psychiatric disorders, and might arise from the disruption in the neural circuits investigated in the present study (30, 31).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by the grants from the National Institute of Health (R01 MH 073246, R01 DA 029330, and T32 NS007224). All primary behavioral and neurophysiological data are archived in the Department of Neurobiology at Yale University School of Medicine.

References and Notes

1. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning. Springer; New York: 2001.
2. Gigerenzer G, Brighton H. Homo heuristicus: why biased minds make better inferences. Top Cogn Sci. 2009; 1:107–143. [PubMed: 25164802]
3. Sutton, RS.; Barto, AG. Reinforcement Learning: An Introduction. MIT Press; Cambridge, MA: 1998.
4. Ito M, Doya K. Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. Curr Opin Neurobiol. 2011; 21:368–373. [PubMed: 21531544]
5. Lee D, Seo H, Jung MW. Neural basis of reinforcement learning and decision making. Annu Rev Neurosci. 2012; 35:287–308. [PubMed: 22462543]
6. Camerer, CF. Behavioral Game Theory. Princeton University Press; New Jersey: 2003.
7. Gallagher HL, Frith CD. Functional imaging of 'theory of mind'. Trends Cogn Sci. 2003; 7:77–83. [PubMed: 12584026]
8. Hampton AN, Bossaerts P, O'Doherty JP. Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc Natl Acad Sci USA. 2008; 105:6741–6746. [PubMed: 18427116]
9. Behrens TE, Hunt LT, Rushworth MF. The computation of social behavior. Science. 2009; 324:1160–1164. [PubMed: 19478175]
10. Zhu L, Mathewson KE, Hsu M. Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning. Proc Natl Acad Sci USA. 2012; 109:1419–1424. [PubMed: 22307594]
11. Suzuki S, et al. Learning to simulate others' decisions. Neuron. 2012; 74:1125–1137. [PubMed: 22726841]
12. Seo H, Lee D. Behavioral and neural changes after gains and losses of conditioned reinforcers. J Neurosci. 2009; 29:3627–3641. [PubMed: 19295166]
13. Nash JF. Equilibrium points in *n*-person games. Proc Natl Acad Sci USA. 1950; 36:48–49. [PubMed: 16588946]
14. Materials and methods are available as supplementary material on *Science* Online.
15. Genovesio A, Brasted PJ, Wise SP. Representation of future and previous spatial goals by separate neural populations in prefrontal cortex. J Neurosci. 2006; 26:7305–7316. [PubMed: 16822988]
16. 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]
17. Tsujimoto S, Genovesio A, Wise SP. Evaluating self-generated decisions in frontal pole cortex of monkeys. Nat Neurosci. 2010; 13:120–126. [PubMed: 19966838]

18. 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]
19. Vickery TJ, Chun MM, Lee D. Ubiquity and specificity of reinforcement signals throughout the human brain. Neuron. 2011; 72:166–177. [PubMed: 21982377]
20. Kim H, Lee D, Jung MW. Signals for previous goal choice persist in the dorsomedial, but not dorsolateral striatum of rats. J Neurosci. 2013; 33:52–63. [PubMed: 23283321]
21. Barraclough DJ, Conroy ML, Lee D. Prefrontal cortex and decision making in a mixed-strategy game. Nat Neurosci. 2004; 7:404–410. [PubMed: 15004564]
22. Seo H, Barraclough DJ, Lee D. Lateral intraparietal cortex and reinforcement learning du ring a mixed-strategy game. J Neurosci. 2009; 29:7278–7289. [PubMed: 19494150]
23. Donahue CH, Seo H, Lee D. Cortical signals for rewarded actions and strategic exploration. Neuron. 2013; 80:223–234. [PubMed: 24012280]
24. Shima K, Mushiake H, Saito N, Tanji J. Role for cells in the presupplementary motor area in updating motor plans. Proc Natl Acad Sci USA. 1996; 93:8694–8698. [PubMed: 8710933]
25. Isoda M, Hikosaka O. Switching from automatic to controlled action by monkey medial frontal cortex. Nat Neurosci. 2007; 10:240–248. [PubMed: 17237780]
26. Hikosaka O, Isoda M. Switching from automatic to controlled behavior: cortico-basal ganglia mechanisms. Trends Cogn Sci. 2010; 14:154–161. [PubMed: 20181509]
27. 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]
28. 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]
29. Botvinick MM. Hierarchical reinforcement learning and decision making. Curr Opin Neurobiol. 2012; 22:956–962. [PubMed: 22695048]
30. Maia TV, Frank MJ. From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci. 2011; 14:154–162. [PubMed: 21270784]
31. Lee D. Decision making: from neuroscience to psychiatry. Neuron. 2013; 78:233–248. [PubMed: 23622061]

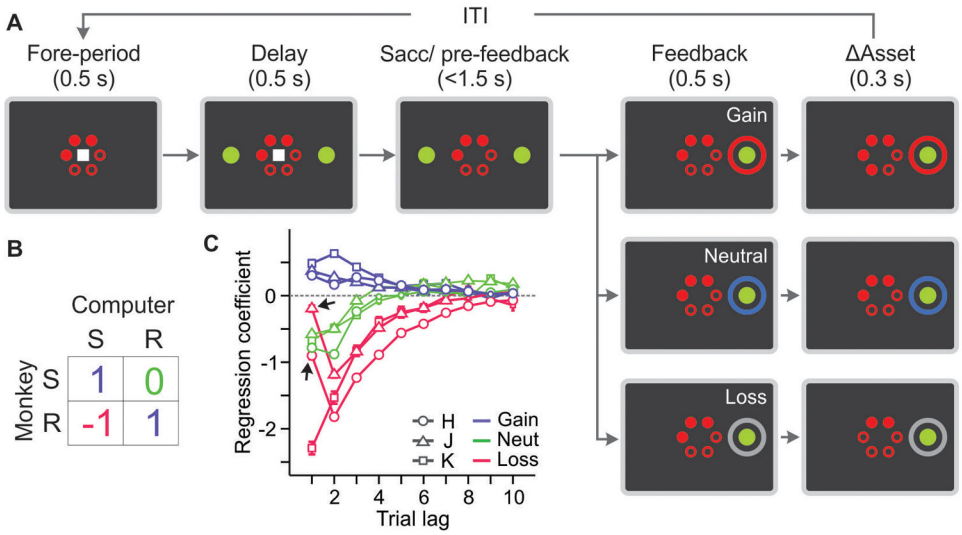

Fig. 1. Behavioral task and performance

Biased matching pennies (A) and its payoff matrix (B). R, risky target; S, safe target. (C) Behavioral effects of gains and losses. Average regression coefficients (ordinate) quantified the tendency for the animal to choose the same target that produced a particular outcome in each of the last 10 trials. Arrows indicate the attenuation in the immediate effect of loss. Error bars, SEM.

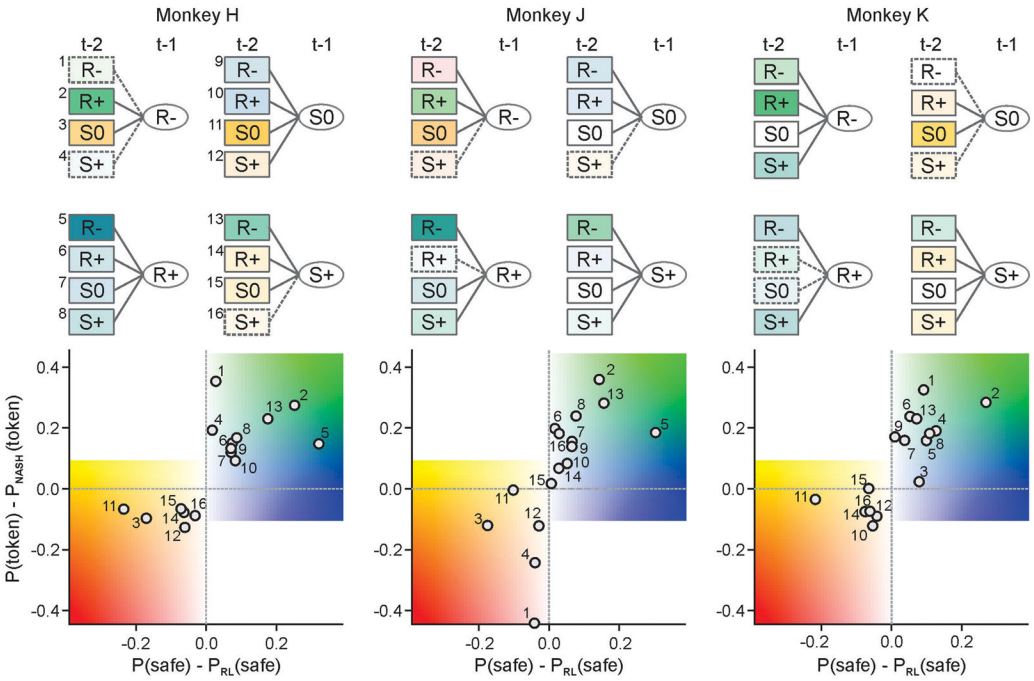

Fig. 2. Systematic deviations from reinforcement learning was beneficial

The color of each box in the decision trees (top) and the position of each circle in the scatter plots (bottom) indicate how much the probability of choosing the safe target deviated from the prediction of the best fitting reinforcement learning model (abscissa in the bottom scatter plot) according to the choices and outcomes in the last two trials, and how this increased or decreased the probability of token compared to the Nash-equilibrium strategy (ordinate in the scatter plot). Numbers indicate different sequences of choices and outcomes in the two preceding trials. Solid boxes correspond to the sequences included in the best hybrid reinforcement learning model (14). R−, and R+ denote loss and gain from the risky target, respectively, whereas S0 and S+ neutral outcome and gain from the safe target.

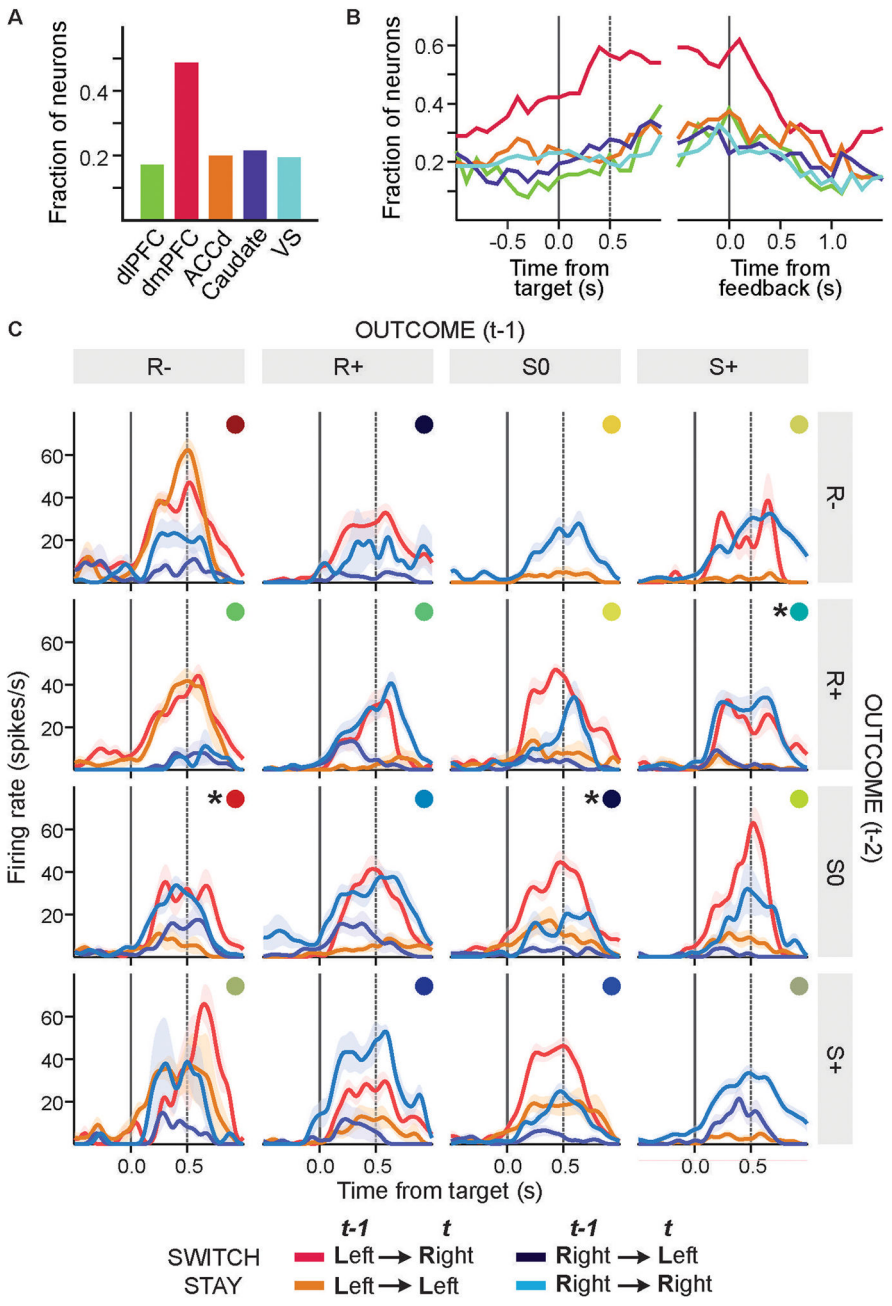

Fig. 3. Cortical activity related to the conjunctions of choices and outcomes

(A) Fraction of neurons in each brain region that significantly modulated their activity during the delay period according to high-order conjunctions of choices and outcomes (14). (B) The time course of signals plotted in (A), using the same color code used to indicate different brain areas. (C) Spike density functions of an example dmPFC neuron sorted by the animal's choices (R, risky; S, safe) and outcomes (+, 0, and − for gain, neutral and loss) as well as the positions of the chosen target in the current (t) and last (t−1) trials. Colored disks indicate different sequences of previous choices and outcomes, and asterisks indicate the activity re-plotted in Fig. 4.

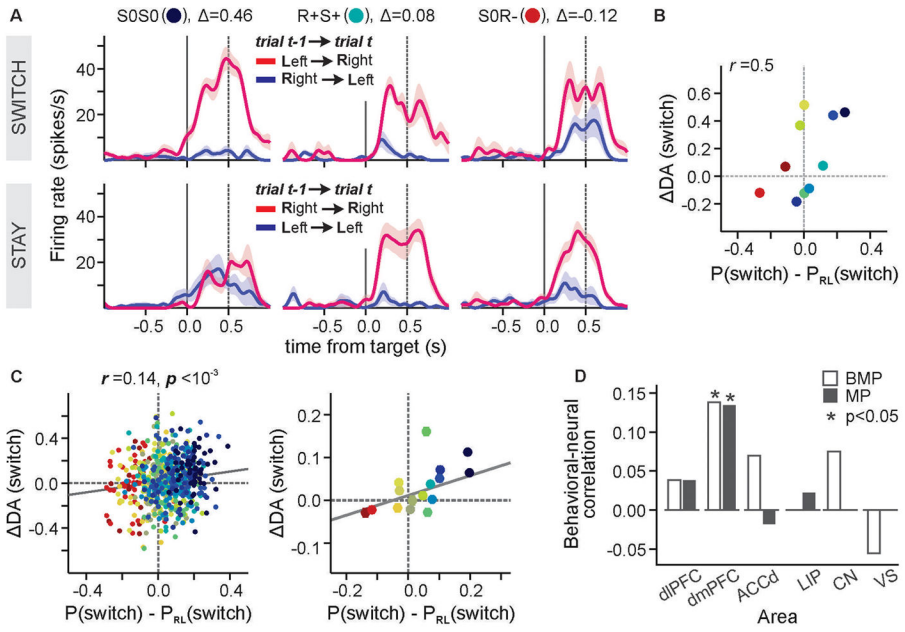

Fig. 4. Cortical signals for deviation from simple reinforcement learning

(A) Spike density functions from a dmPFC neuron (shown in Fig. 3C) sorted by the animal's choices in the current and previous trials for 3 different sequences of outcomes in the last two trials (indicated by the text label and color defined in Fig. 3C). Δ denotes the difference in the accuracy of decoding the animal's choice in switch vs. stay trials. (B) The difference in the decoding accuracy, ΔDA(switch), plotted as a function of how much more often the animal switched its choices compared to the prediction from the simple RL algorithm. (C) The same results shown in (B) for the entire population of dmPFC neurons (left) and averaged for each outcome sequence (identified by colors defined in Fig. 3C; right). Lines correspond to the best-fitting regression models. (D) The correlation coefficient between ΔDA and the deviation from reinforcement learning model for two different data sets (BMP, biased matching pennies; MP, matching pennies).