Behavioral and neural evidence for dynamic model arbitration in dorsolateral prefrontal cortex

Jae Hyung Woo (jae.hyung.woo.gr@dartmouth.edu)*
Department of Psychological and Brain Sciences, Dartmouth College
Hanover, NH 03755, USA

Michael Chong Wang (chong.wang.gr@dartmouth.edu)*
Department of Psychological and Brain Sciences, Dartmouth College
Hanover, NH 03755, USA

Ramon Bartolo (ramon.bartoloorozco@nih.gov)
Laboratory of Neuropsychology, National Institute of Mental Health
Bethesda, MD 20892, USA

Bruno B. Averbeck (averbeckbb@mail.nih.gov)
Laboratory of Neuropsychology, National Institute of Mental Health
Bethesda, MD 20892, USA

Alireza Soltani (alireza.soltani@dartmouth.edu)
Department of Psychological and Brain Sciences, Dartmouth College
Hanover, NH 03755, USA

* Equal contribution

Abstract

One of the hallmarks of higher cognitive function is the ability to link outcomes to relevant features of the environment, while ignoring the irrelevant features. This is especially relevant for learning and decision making under uncertainty, where any features or attributes of a selected option can be predictive of rewards. It has been suggested that the brain tackles such uncertainty by running multiple internal models of the environment and arbitrating among them based on their reliability. To reveal the neural mechanisms underlying this dynamic arbitration process, we carried out high channel count recordings in dorsolateral prefrontal cortex (dlPFC) while monkeys performed a probabilistic reversal learning task with multiple layers of uncertainty. By fitting choice behavior with models based on reinforcement learning, we found evidence for dynamic, competitive interaction between stimulus-based and action-based learning strategies. dlPFC was involved in arbitration in two ways: (1) arbitration weight was represented in the activity of dlPFC neurons; (2) only the relevant information for the currently adopted strategy was encoded congruently as the monkey's subsequent choice. These results suggest that dlPFC could be crucial for flexible arbitration between alternative models of the environment.

Keywords: decision-making, reinforcement learning; cognitive control; dorsolateral prefrontal cortex

Introduction

One of the most challenging aspects of learning in naturalistic settings is that it is inherently uncertain which features of the environment are predictive of rewards. To form an appropriate decision, the brain must select and learn from only the relevant features of a choice option or a preceding action, while suppressing signals from irrelevant ones. It has been suggested that the brain tackles such uncertainty by running multiple internal models of the environment, each predicting outcomes based on different attributes of choice options, and using the reliability of these predictions to select the appropriate model to inform choice behavior (Soltani & Koechlin, 2022; Averbeck & O'Doherty, 2022).

To reveal the neural mechanisms underlying this arbitration, we studied the choice behavior of monkeys performing a probabilistic reversal learning task with uncertainty about the correct model of the environment. We constructed multiple models based on reinforcement learning (RL) to fit choice behavior on a trial-by-trial basis. We also investigated neural signals related to arbitration in dlPFC, which is known to encode both task-relevant and irrelevant variables (Seo et al., 2007; Donahue et al., 2013; Donahue & Lee, 2015; Tsutsui et al., 2016) as well as inference about the current state of the environment (Genovesio et al., 2005; Bartolo & Averbeck, 2020).

Methods

Two rhesus monkeys performed a variant of two-armed bandit task. On a reversal trial, the reward probability for the better and worse option (70/30%) flipped (**Fig.1A**). Critically, unbeknownst to the monkeys, the reward assignment for a given block was either stimulus- (*What*) or action-based (*Where*). In What blocks, rewards were assigned based on stimulus identity. In Where blocks, rewards were assigned based on the chosen location, regardless of objects appearing on that side. Two block types were randomly interleaved throughout the session. Neural population activity was recorded with eight Utah arrays implanted bilaterally in area 46 (**Fig.1B**), resulting in a total of 6132 recorded cells across eight recording sessions.

Analysis of behavioral and neural data

To estimate the strategy used for each block, we ran logistic regression on choice behavior. To reveal neural mechanism underlying arbitration, we also tested multiple reinforcement learning (RL) models and used five-fold cross validation to compare goodness of fit.

Two-system RL with reliability-based arbitration. We considered a hybrid RL model consisting of two value functions, V_{Stim} and V_{Action}, to simultaneously track the values of stimuli and actions. Decision value (DV) for each side (left or right) was computed by combining two value functions with arbitration weight ω as follows:

$$DV_i = V_{Stim(i)}\omega + V_{Action(i)}(1-\omega),$$

where $V_{Stim(i)}$ indicates the value of stimulus appearing on side *i* (left or right) for a given trial. At the end of every trial, ω values (ranging [0, 1]) were updated as:

$$\omega(t+1) = \omega(t) + \psi|\Delta Rel|\big(I - \omega(t)\big) + \xi\big(\omega_0 - \omega(t)\big),$$

where $\Delta Rel = V_{C,Stim}(t) - V_{C,Action}(t)$ is the difference in reliability between two systems on each trial, defined as the value of chosen options (V_C) in each system. ψ is the arbitration rate, I indexes update direction for ω (1 if ΔRel>0, or 0 otherwise), ω_0 is the initial ω on the first trial of a block, and ξ is the decay rate in ω toward its initial value. An alternate definition of reliability signal with $\Delta Rel = |RPE_{Action}(t)| - |RPE_{Stim}(t)|$ was also tested. We also considered an alternate model without dynamic arbitration, fitting a fixed value of ω for every block. All models included separate learning rates for rewarded and unrewarded trials, decay for unchosen option, an inverse temperature, and a side bias term.

Linear regression analysis. We used a linear regression model to investigate single-unit activities in dlPFC, in 50ms bins time-locked to cue onset. The predictors included: currently/previously chosen image or location, current/previous reward outcomes, previously rewarded image or location, current position of images, and location of previously chosen or rewarded image. These terms were nested within the adopted strategy type, consisting of three levels ("action-dominant," "mixed," "stimulus-dominant") as inferred through tertile split of dynamic ω values from the RL model. Namely,

trials with lower ω values toward 0 were categorized as "action-dominant", while those with higher ω toward 1 were categorized as "stimulus-dominant." This allowed us to study the relationship among the regressors by monkey's adopted strategy. The regression also included the main effects of ω and its distance from maximum uncertainty (|ω-0.5|) as continuous predictors.

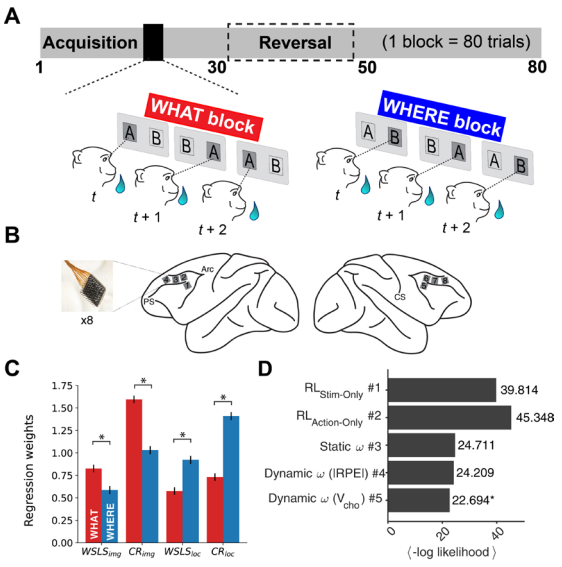

Figure 1: (**A**) Schematic of the task and different block types. (**B**) Location of eight microelectrodes arrays. (**C**) Coefficients for logistic regression by block types. (**D**) Five-fold cross-validation of RL models.

Results & Discussion

We first examined logistic regression coefficients for win-stay/lose-switch (WSLS) and choice repetition (CR) terms on either chosen image or location (**Fig.1C**). We found that WSLS and CR for images were larger during What blocks, while those for location were larger during Where blocks (two-sample t-test, P<.001). That is, monkeys overall adopted appropriate stimulus-based (action-based) strategy for What (Where) blocks.

Next, we compared the fit of choice behavior among the RL models. Mean negative log-likelihoods from 100 unique cross-validation instances are shown in **Fig.1D**. Note that the baseline models that only learn the values of either stimuli (#1) or actions (#2) were insufficient to explain monkeys' choice behavior. In comparison, the two-system RL with fixed arbitration weight (#3) significantly improved the fit. Adding dynamic arbitration component to the model further improved the fit. Importantly, reliability signal based on the value of chosen option (#5) better captured the arbitration process than that based on the magnitude of reward prediction error (#4).

Using the estimated dynamic arbitration weights ω from the best model (#5), we next examined the regression on dlPFC activities. Critically, we found evidence for encoding of ω values (and also |ω-0.5|) significantly above the chance level of 5% (binomial test, P<.001; **Fig.2A**). Population activities also reflected encoding of other task-related variables from the current trial (**Fig.2B**), previous trial (**Fig.2C**), and their interactions (**Fig.2D**).

To examine the neural signature of arbitration and its effect on choice behavior, we next observed the relationship among regressor coefficients (200ms after cue onset) by adopted strategies. We found that regression coefficients for previous and current chosen location were significantly correlated for action-dominant (Spearman's, P<.001; **Fig.2E**, blue) but not for stimulus-dominant trials (**Fig.2E**, red). Conversely, regressor for the location of previously chosen image was highly correlated with chosen location during stimulus-dominant trials (**Fig.2F**, red). In other words, neurons that increased activity when previously chosen image was on the right also tended to increase their activities when rightward choice was made. In contrast, the two regressors were negatively and weakly correlated during action-dominant trials (**Fig.2F**, blue). These results suggest that only the relevant information for the currently adopted strategy was encoded in the same population subspace as the monkey's subsequent choice.

Together, our results illustrate behavioral and neural signatures of dynamic arbitration between stimulus- and action-based strategies. In particular, dlPFC neurons directly encoded arbitration weight ω, and showed aligned encoding of choice according to the current strategy in use as signaled by ω. Thus, dlPFC may be critical for flexible switching between competing models of the environment.

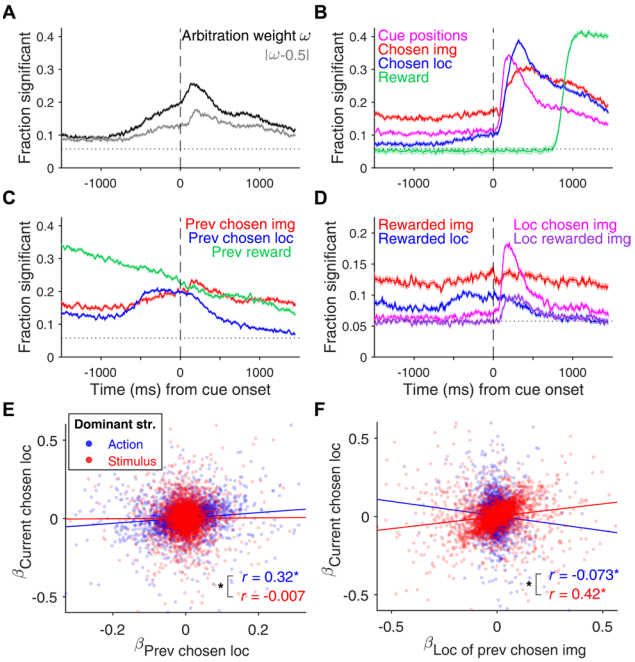

Figure 2: (**A-D**) Fraction of significant neurons (P<.05) for each regressor. (**E-F**) Congruent coding of current choice location with previous choice (E) or location of previously chosen image (F) by dominant strategy.

Acknowledgments

This work is supported by the National Institutes of Health (R01 DA047870 to A.S.) and by Intramural Research Program of the NIMH (ZIA MH002928).

References

Averbeck, B., & O'Doherty, J. P. (2022). Reinforcement-learning in fronto-striatal circuits. *Neuropsychopharmacology*, *47*(1), 147-162.

Bartolo, R., & Averbeck, B. B. (2020). Prefrontal cortex predicts state switches during reversal learning. *Neuron*, *106*(6), 1044-1054.

Donahue, C. H., & Lee, D. (2015). Dynamic routing of task-relevant signals for decision making in dorsolateral prefrontal cortex. *Nature neuroscience*, *18*(2), 295-301.

Donahue, C. H., Seo, H., & Lee, D. (2013). Cortical signals for rewarded actions and strategic exploration. *Neuron*, *80*(1), 223-234.

Genovesio, A., Brasted, P. J., Mitz, A. R., & Wise, S. P. (2005). Prefrontal cortex activity related to abstract response strategies. *Neuron*, *47*(2), 307-320.

Seo, H., Barraclough, D. J., & Lee, D. (2007). Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. *Cerebral Cortex*, *17*(suppl_1), i110-i117.

Soltani, A., & Koechlin, E. (2022). Computational models of adaptive behavior and prefrontal cortex. *Neuropsychopharmacology*, *47*(1), 58-71.

Tsutsui, K. I., Grabenhorst, F., Kobayashi, S., & Schultz, W. (2016). A dynamic code for economic object valuation in prefrontal cortex neurons. *Nature communications*, *7*(1), 12554.