Prefrontal cortex predicts state switches during reversal learning

Authors:
Ramon Bartolo and Bruno B. Averbeck

Supplementary Figures

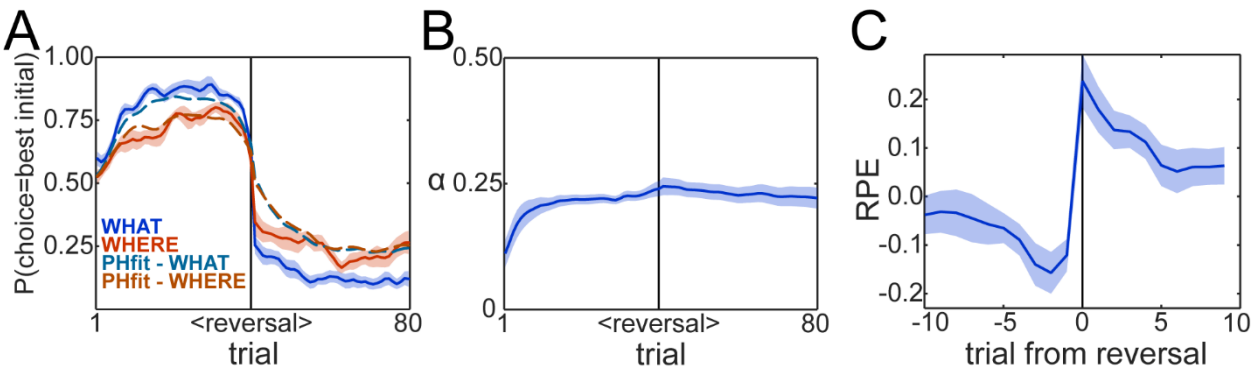

Figure S1. Pearce-Hall reinforcement learning model fitting. Related to Figure 1. **A.** Choice and model data aligned to the BHV reversal estimate. Plots show the fraction of times the animals chose the option that had a higher reward probability at the beginning of the block split by block type. Overlays are choice probability estimates derived from Pearce-Hall model fittings to the choice data. **B.** Associability parameter (α) estimates over trials from the Pearce-Hall model fittings, pooled for both block types. **C.** Reward Prediction Errors from the Pearce-Hall fittings. Data are mean±SEM across sessions (n=8).

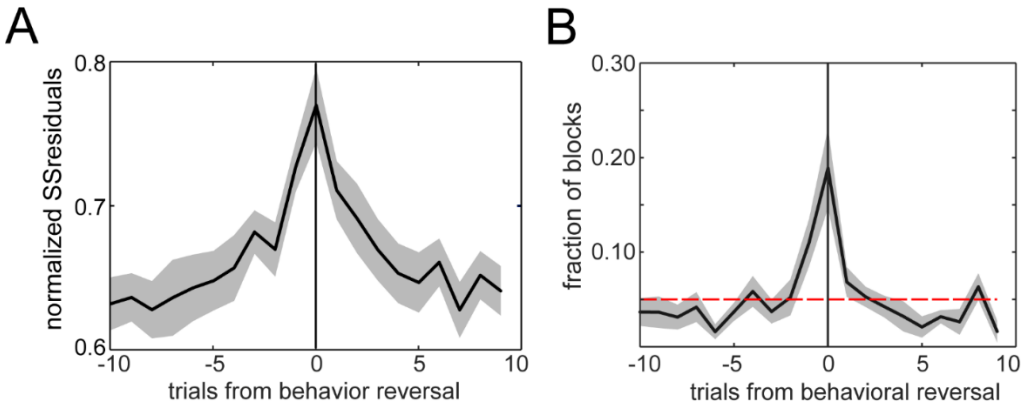

Figure S2. Decoding of Reversal at cue onset from *SSresid*. Related to Figure 3. **A.** Sum of Squared Residuals across neurons. **B.** Histogram of decoded trial of reversal. Within the switch window, we searched for the trial with the maximum *SSresid* and considered this trial as the predicted reversal. Decoded reversals are labeled as the number of trials from the Bayesian point estimate for the behavioral reversal. The red dashed line shows chance level. Values are means±SEM across sessions (n=8).

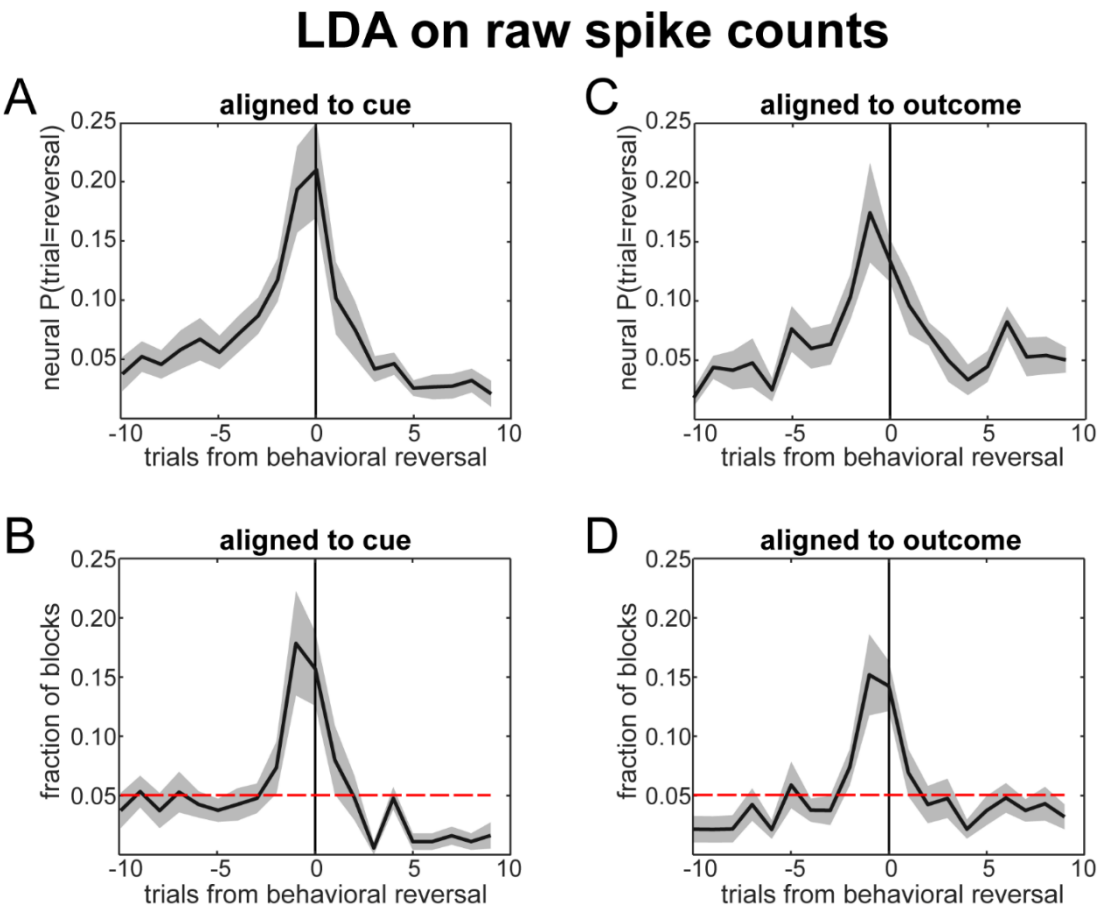

Figure S3. Decoding of Reversal from raw spike count data. Related to Figure 3. **A.** Posterior *P(trial=reversal | neural response)* using spike counts in a window from 0-300ms from cue onset. **B.** Distribution of decoded trial of reversal using cue aligned spike counts. **C-D.** Same as **A-B** but using spike counts in a window from 0-300ms from trial outcome. Trials are labeled as the number of trials from the Bayesian point estimate for the behavioral reversal. The red dashed lines in **B** and **D** show chance level. Values are means±SEM across sessions (n=8).

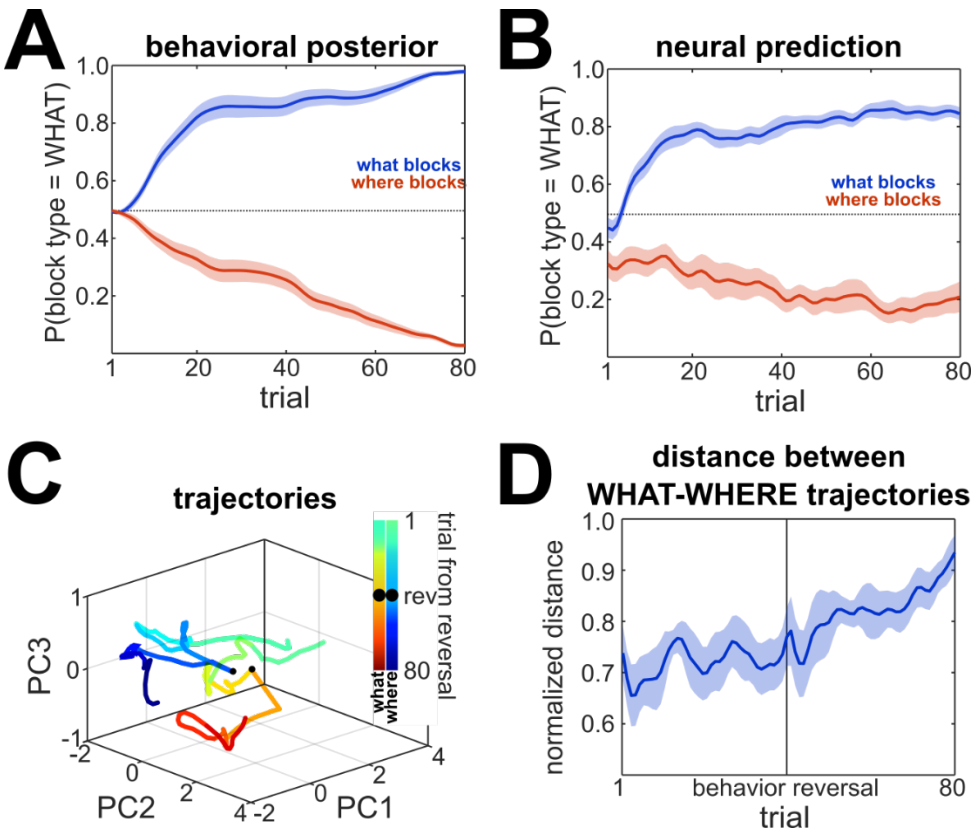

Figure S4. Decoding Block Type from raw spike count data. Related to Figures 6 and 7. **A.** Posterior *P(block type=WHAT | M=BHV)*. **B.** Predicted *P(block type=What | neural response)* using multiple linear regression model, regularized with early stopping, that used spike counts in a window from 0-300ms from cue onset to fit the Bayesian estimated Block Type shown in panel A. **C.** Neural trajectories for an example recording session across block execution, as in main Figure 7A but here split by block type. Trial number in block is color coded and the trial of reversal is indicated by a black dot on each. **D.** Euclidean distance (normalized by the maximum observer value on each session) between the trajectories for WHAT and WHERE blocks. Data are means±SEM across sessions (n=8).