Orbitofrontal neuron ensembles contribute to inhibitory control

Pragathi Priyadharsini Balasubramani¹ & Benjamin Y. Hayden²

1. Department of Psychiatry
University of California San Diego
La Jolla, CA. 92037
2. Neuroscience and Center for Magnetic Resonance Research
University of Minnesota
Minneapolis, MN. 55455

Corresponding Author:

Pragathi Priyadharsini Balasubramani
Department of Psychiatry
University of California San Diego
La Jolla, CA. 92037
Email: pragathipriyadharsini@gmail.com

SUMMARY

Stopping, or inhibition, is a form of self-control that is a core part of adaptive behavior. We hypothesize that inhibition commands originate, in part, from the orbitofrontal cortex (OFC). We recorded activity of OFC neurons in macaques performing a stop signal task. Decoding analyses revealed a clear difference in ensemble responses that distinguish successful from failed inhibition that begins after the stop signal and before the stop signal reaction time. We also found a different and unrelated ensemble pattern that distinguishes successful from failed stopping before the beginning of the trial. These signals were distinct from, and orthogonal to, value encoding, which was also observed in these neurons. The timing of the early and late signals was, respectively, consistent with the idea that OFC contributes both proactively and reactively to inhibition. These results support the view, inspired by anatomy, that OFC gathers diverse sensory inputs to compute early-stage executive signals.

INTRODUCTION

Inhibition is a key function of the brain's executive control system (Aron et al., 2015; Hampshire and Sharp, 2015; Logan et al., 2015). The process of inhibitory control over a prepotent behavior is often studied with the stop signal task, in which the subject must deliberately withhold a planned action in response to a specific stop-related signal (Eagle et al., 2007; Li et al., 2006; Logan, 1994; Schall, 2001). Performance in this task predicts dysregulated executive function in psychiatric conditions like drug addiction, obsessive compulsive disorder, and obesity (e.g Chamberlain et al., 2006; Iacono et al., 2008; Logan et al., 1997; Nederkoorn et al., 2006; Schachar et al., 1995). One convenient feature of the task is that it is validated in several species, meaning that animal models can be used to provide insight into the mechanisms of human inhibition (Hanes and Carpenter, 1999; Hanes and Schall, 1995; Logan and Irwin, 2000; See Eagle et al., 2008; Pouget et al., 2017, for review). Understanding those mechanisms holds promise in developing rational treatments for psychiatric diseases and may also help address philosophical questions about self-control and the nature of volition (Schall et al., 2002).

In order to stop effectively, our brains must monitor both the sensory world and the internal milieu for information indicating that planned actions have become disadvantageous and should be cancelled. When this control is directly driven by external signals, such as a stop signal, it is known as reactive control (Braver, 2012; Braver et al., 2007; Chen et al., 2010). In the stop signal task (Logan, 1994; Logan and Cowan, 1984), reactive control can be identified because it occurs after the presentation of a stop signal and before the inferred behavioral response to it, the stop signal reaction time (SSRT, See

1 Schall, 2001; Schall et al., 2002 for review). Stopping decisions can also be influenced
2 and in some cases are entirely determined by internal processes, which can begin even
3 before the start of a trial. This form of inhibitory control is known as proactive control
4 (Braver, 2012; Braver et al., 2007; Chen et al., 2010).

5 Signatures of reactive control have been observed in the frontal eye fields (FEF)
6 and the primary motor cortex, as well as in midbrain structures like the superior
7 colliculus (SC). Many medial prefrontal structures such as supplementary motor area
8 (SMA), pre-SMA, and SEF show signatures of proactive control (Chen et al., 2010;
9 Chikazoe et al., 2009; Majid et al., 2013; Stuphorn et al., 2010; Stuphorn and Emeric,
10 2012). In the case of eye movements, control is most directly determined by processes
11 occurring in the FEF and SC. In these regions, inhibition is driven by a rapid rise in firing
12 rates of a specific subpopulation of neurons—fixation neurons - that gate the activity of
13 another subpopulation—movement neurons (Hanes and Schall, 1995; Logan et al., 2015;
14 Schall, 1991).

15 What is the source of these inhibition signals? In our view, inhibition likely does
16 not emerge from a single specialized stopping area, but rather reflects the integration of
17 diverse forms of information, at varying levels of abstraction, bearing on the need to stop
18 (Hampshire and Sharp, 2015). Such pre-inhibition signals are likely to be especially
19 prominent throughout the prefrontal cortex, which, directly and indirectly, is positioned
20 to regulate motor processes (Aron, 2007; Duncan, 2001; Hampshire and Sharp, 2015;
21 MacLeod et al., 2003). We were particularly interested in the orbitofrontal cortex (OFC),
22 a region on the orbital surface that is closely associated with value and reward processing
23 (Padoa-Schioppa, 2011; Rudebeck and Murray, 2014; Schoenbaum et al., 2009; Wallis,

2007). The OFC is the major input for sensory information into the PFC: it receives strong visual, auditory, gustatory, and olfactory inputs. It also has access to signals relating to internal states, via the medial network (Öngür and Price, 2000). It is also often proposed to occupy an early position in PFC processing hierarchies (Carmichael and Price, 1994, 1996; Fuster, 1988; Fuster, 2001; Rushworth et al., 2012; Rushworth et al., 2011). These facts raise the possibility that OFC serves as a first stage (or at least a relatively early stage) for computing preliminary executive signals including, potentially, inhibition ones.

The contribution of the OFC to inhibition remains disputed. On one hand, multiple studies give the OFC a prominent role in inhibition (Bryden and Roesch, 2015; Chikazoe et al., 2009; Dias et al., 1996; Eagle et al., 2007; Horn et al., 2003; Majid et al., 2013; Mishkin, 1964; Roberts and Wallis, 2000). On the other, a good deal of work argues against such a direct inhibitory role of OFC (Chudasama et al., 2006; Ghods-Sharifi et al., 2008; Rudebeck and Murray, 2014; Schoenbaum et al., 2003). Complementary work suggests that inhibition may be the specific domain of functionally and anatomically distinct structures, especially ones located on the ventrolateral surface (Aron et al., 2015; Aron et al., 2003; Banich and Depue, 2015). Finally, another current of research sees the function of the OFC as purely or primarily economic, and thus unlikely to play any major role in inhibition per se (Padoa-Schioppa, 2011; Wallis, 2007). However, recent work indicates that the OFC may have a clear executive role above and beyond its economic one (Bryden and Roesch, 2015; Mansouri et al., 2014; Mansouri et al., 2009; Sleezer et al., 2016; Sleezer et al., 2017). This executive role may include a contribution to inhibition among other functions. Supporting this hypothesis, a recent

1 study found greater spatial selectivity in rodent OFC on successful inhibition trials than
2 on failed inhibition trials, indicating that OFC carries executive signals that directly
3 regulate inhibitory processes (Bryden and Roesch, 2015).

4 We hypothesized that OFC participates in stopping decisions by modulating
5 responses in downstream regions that in turn regulate motor activity. In an oculomotor
6 task, such signals would have to have appropriate timing to influence stopping. Thus, we
7 predict that successful vs. failed stopping would be correlated with different neuronal
8 ensemble patterns with specific timecourses. We therefore examined ensemble states of
9 OFC neurons recorded in a stop signal task. We found a significant coding pattern
10 difference that emerged following the stop signal but before the stop signal reaction time.
11 We also found a distinct (i.e. statistically orthogonal) ensemble pattern difference that
12 was observable before the trial onset and that derived from the same set of neurons. Both
13 of these patterns were distinct from (also orthogonal to) economic (i.e. value) signals,
14 which were also carried by the same neurons. Together, these pattern differences provide
15 evidence that OFC carries information sufficient to influence inhibition and suggest it
16 may do so both reactively and proactively.

17

18

METHODS

Subjects

Two male rhesus macaques (*Macaca mulatta*, subject J and subject T) served as subjects. All animal procedures were approved by the University Committee on Animal Resources at the University of Rochester and were designed and conducted in compliance with the Public Health Service's Guide for the Care and Use of Animals.

Recording site

A Cilux recording chamber (Crist Instruments) was placed over the area 13 of OFC (*Figure 1B*). The targeted area expands along the coronal planes situated between 28.65 and 33.60 mm rostral to the interaural plane with varying depth. Position was verified by magnetic resonance imaging with the aid of a Brainsight system (Rogue Research Inc). Neuroimaging was performed at the Rochester Center for Brain Imaging, on a Siemens 3T MAGNETOM Trio Tim using 0.5 mm voxels. We confirmed recording locations by listening for characteristic sounds of white and grey matter during recording, which in all cases matched the loci indicated by the Brainsight system.

Electrophysiological techniques

Single electrodes (Frederick Haer & Co., impedance range 0.8–4 MOhm) were lowered using a microdrive (NAN Instruments) until waveforms of between one and five neuron(s) were isolated. Individual action potentials were isolated on a Plexon system. Neurons were selected for study solely based on the quality of isolation; we never preselected based on task-related response properties.

Eye tracking and reward delivery

Eye position was sampled at 1,000 Hz by an infrared eye-monitoring camera system (SR Research). Stimuli were controlled by a computer running MATLAB (Mathworks) with Psychtoolbox (Brainard and Vision, 1997) and Eyelink Toolbox (Cornelissen et al., 2002). A standard solenoid valve controlled the duration of water delivery. The relationship between solenoid open time and water volume was established and confirmed before, during, and after recording.

Task paradigm

The task followed standard stop signal task paradigm (Logan, 1994; Logan and Cowan, 1984). Subjects were placed in front of a computer monitor (1920x1080px) with black background. Following a brief (300 msec) central fixation on a white circle (radius 25px, *Figure 1*), the fixation spot disappeared on the appearance of eccentric saccade target (90px white square, 2.38 degrees, positioned at 288px in left or 1632px in right of screen, 50% chance). A go trial (67% of trials, randomly selected) was indicated by a go signal which is the peripheral target, whereas a stop trial (33% of trials, randomly selected) was indicated by an additional appearance of a stop signal—a central gray

square (90px square, 2.38 degrees) delayed relative to the go signal presentation. Stop signal delays (SSD) in the task were set to stabilize at a delay causing approximately 50% successful stopping out of all stop trials recorded for the task in that day; SSDs were modulated through a staircase procedure with intervals of 16 msec. On go trials, subjects were rewarded for a saccade to the go signal and fixating on it for 200 msec; and on stop trials, subjects were rewarded for inhibiting their saccade to go signal and fixating at the stop signal for 400 msec. Water rewards were provided as feedback, and they were contingent on subject's performance. Rewards were always 125 μ l. The inter trial interval was 800 msec.

The *economic choice task* had a similar task framework to stop signal task, and they interleaved randomly in an interval of 1-3 trials. In go trials (random 67% of the total), a peripheral target called go offer (90px white square, 2.38 degrees, positioned at 288px in left or 1632px in right of the screen, 50% chance) was presented, and it was randomly associated with low (15 μ l), medium (125 μ l), or high (250 μ l) reward offers, as indicated by yellow, blue and magenta colored squares, respectively. In stop trials (random 33% of the total), a center stop offer (90px square, 2.38 degrees) delayed with respect to the appearance of go offer was presented in addition. The stop-offer was also randomly associated with yellow, blue and magenta colors to indicate low, medium and high reward sizes. The go offer in stop trials was always in blue color to represent medium reward sized offer. This setup allowed the subject to make a choice through reward comparison in case of stop trials, and through a forced choice in case of go trials. All other parameters were the same as stop signal task.

Behavioral analysis

Inhibition function related failed inhibitions to stop signal delay (SSD). The delay from the presentation of go signal that caused 50% successful cancellation in stop signal task (SSD-50) was used for computing stop signal reaction time (SSRT). SSRT was usually computed through median and integration methods (Logan, 1994; Logan and Cowan, 1984; Verbruggen and Logan, 2008). *Median* method computed median of go trials' reaction time distribution and then subtracted SSD-50 from it to give SSRT. The *integration* method computed the point in go trials' RT distribution whose area was half the whole and then subtracted SSD-50 from it to give SSRT. SSRT computed from both of the above methods gave nearly equal results, and they were averaged to obtain the final SSRT estimates reported for both subjects.

Statistical methods

Separate PSTH matrices were constructed by aligning spike rasters to the presentation of the go signal and stop signal for every neuron. Firing rates were calculated in 1 msec bins but were generally analysed in longer epochs. Normalization procedure was carried out by subtracting the mean firing during inter-trial interval (ITI) time period and then by zscoring each neuron's data. For display, PSTHs were smoothed using 200 msec running boxcars. Tests used in the study include two sample t-test for parameteric analysis, Wilcoxin rank test for non-parametric analysis, chi-square test for comparing decoder's classification accuracy against baseline (50% classification

accuracy), Pearson correlation method for correlation analysis. To compute population tuning, we picked neurons with significant ($p < 0.05$) differences between successful and failed inhibition trials using Wilcoxin rank test.

Decoding analyses

We chose a neural network based decoding technique because it could efficiently analyse population responses in frontal cortex that are highly multiplexed and non-linear. To generate population activation states as input patterns for the decoding analysis, we first separated all trials of each neuron by trial conditions (successful and failed inhibition trials). Then, we averaged the activity from randomly sampled 10 trials belonging to a condition, with replacement, to form activation state for a neuron in any particular time period. The averaged responses of all 96 neurons' were pooled to generate one population activation state for a particular trial condition and for a specific time period. 100 unique activation patterns were used for the network training. The procedure is similar to that carried out by other studies (e.g., Mante et al., 2013; Pouget et al., 2000; Rigotti et al., 2013; Wang and Hayden, 2017)

The network used to study the stopping patterns had 100 hidden nodes, and 2 output nodes each representing one target condition for classification. The number of input nodes equal to the total number of neurons used for analysis = 96 (from two subjects). The network weights were initialized to small random numbers between -0.01 and 0.01.

The following *back-propagation algorithm* was used for training the decoders (Haykin and Network, 2004; Rumelhart et al., 1986; Werbox, 1974). In the below, the input nodes are denoted by subscript, k , hidden nodes by subscript, j , and output nodes by subscript, i . Output error, e , associated with the network's response for the p 'th input pattern was

$$e_i = \text{desired output} - y_i \quad (1)$$

where y_i was the i 'th output node response, and desired output was 1 / 0 if the i 'th output node was associated with target trial condition for the corresponding input pattern (e.g., successful inhibition, failed inhibition). Total output error over all input patterns was computed by,

$$E = \sum_p E_p \quad (2)$$

Network's objective was to minimize the squared output error (eqn. 1) for the p 'th pattern as denoted by eqn. (3).

$$E_p = \frac{1}{2} \sum_i e_i^2 \quad (3)$$

Response of any node was a hyperbolic tangent function (g) of slope = 5 of the total input (h_i^s) to it. The output node response, y_i , as a function of its input was calculated as,

$$y_i = g(h_i^s) \quad (4)$$

where, net input (h_i^s) to the output layer was,

$$h_i^s = \sum_j w_{ij}^s V_j \quad (5)$$

In the above, the weights, w_{ij} , with superscript, s , indicate the second level of the network between hidden and output layer. V_j denoted the output of hidden layer, and it was represented as a function of net input to the hidden node (h_i^f) as follows,

$$V_j = g(h_j^f) \quad (6)$$

and

$$h_j^f = \sum_k w_{jk}^f x_k \quad (7)$$

The superscript, f , in eqns. (6, 7) denote first level of the network between input and hidden layer, w_{jk} were their weights, and x_k was the input pattern to neural network. Weight updates were proportional to the negative change in error for the p 'th pattern, E_p , on change in weights. All updates happened trial by trial in the training phase. The update used at the second level was by eqn. (8), and that in the first level was by eqn. (10).

$$\Delta w_{ij}^s = -\eta \frac{\partial E_p}{\partial w_{ij}^s} = \eta \delta_i^s V_j \quad (8)$$

where,

$$\delta_i^s = e^{-g'(h_i^s)} \quad (9)$$

$$\Delta w_{jk}^f = \eta \delta_j^f x_k \quad (10)$$

where,

$$\delta_j^f = \sum_i \delta_i^s w_{ij}^s g'(h_j^f) \quad (11)$$

η is the learning rate set to 0.001 for pre-go and post-stop signal decoder, and 0.01 for reaction time decoder, and g' denotes first order derivative of hyperbolic tangent function.

We had two different decoders trained on data from 1) pre-go signal, 2) post-stop signal time periods; the former worked on data aligned to presentation of go signal at time = 0, and the latter worked on data aligned to stop signal. For *pre-go decoder*, the training data was population activation states generated on averaging the signal from the fixation epoch spanning 300 msec before the presentation of go signal. For *post-stop decoder*, training data was generated on averaging the firing between 100 msec to 250 msec of stop signal presentation. The entire network was run for $n = 100$ instances with different random weight initializations to obtain average output performance. Training procedure in all instances converged to classification accuracy of above 80%, and the converged weights at the end of training were used for testing of decoder. The testing data used were population activation states generated by averaging 100 msec boxcars that slides with step size of 10 msec (a total of 91 boxcars). The kind of normalization methods didn't alter the results (See *Figure S2*).

Similarities in the functioning and generalization of pre-go and post-stop decoders were analysed by comparing their converged weights, as well as by comparing their

classification accuracy. The similarity index was computed by cross correlating converged hidden layer weight vectors (with zero lag) of two decoders of interest. The index was averaged across n ($=100$) instances of networks with different weight initializations. The similarity index obtained from autocorrelating the weight vectors were used to statistically compare and cross-validate the results from cross correlation, and the results were significant using ttest (ttest, tstat = 210, $p < 0.001$). Similarities in classification accuracy at pre-go or post-stop signal period were found by using t-test on average performances of the two decoders during n instances (with different random weight initializations).

Cancellation time was defined by the size of test-boxcar window positioned at first instance of atleast four consecutive test-boxcars (100 msec window moving in intervals of 10 ms) in a row, whose performance was significantly higher than 50% using chi-square test ($p < 0.05$). The method avoids false positives that otherwise appear by 99% chance when considering just any one single significant instance of 91 total boxcars. With simulations using markov chains, we found that at least 4 consecutive significant windows were needed in a row for the claim of significance with $p < 0.001$; so the criteria to find at least 4 consecutive significant bins were used to find pre-go and post-stop decoder results (figure 4) as well as cancellation time. Average latency of cancellation signals to SSRT was found by subtracting SSRT of each subject from the mean cancellation time (90 msec).

Incase of the decoder used for analyzing the reaction time ensemble patterns, the inputs to the decoder were either the population activation pattern during time periods 200 msec before or after the reaction time, appropriate to the case of analysis. The output of the decoder was the number of coarse reaction time bins for classifying the input data ($n = 5$) in the range 0.1 – 0.6 sec. The neurons that contain enough data for all reaction time bins were only considered for this analysis. For reaction time bins taken to be a total of 5, we considered 26 neurons for ensemble analysis. A sum of hundred ensemble patterns was generated for training. The training procedure was similar to pre-go and post-stop signal decoders, and it was carried out for 75% of data. 25% of data was used for testing. The results presented were significant when statistically analyzed against chance percent of 20% using chi-square tests. The decoder results were also cross validated with input as random-data to find that the reaction time ensemble pattern classification was statistically significant with $p < 0.001$ (ttest, tstat = 27.91).

Reward and stopping index

Reward index for every neuron was measured by linearly regressing the firing at outcome epoch (between reaction time and feedback) to the received reward sizes in *neuroeconomic trials*. The stopping index was measured as the difference in normalized firing rates (FR) of successful and failed inhibition trials divided by their norm.

$$\text{Stopping index} = \frac{FR_{sc} - FR_{snc}}{\sqrt{FR_{sc}^2 + FR_{snc}^2}} \quad (12)$$

Cross validation tests were performed to support the idea that we had sufficient

1 data to detect an effect had it been there, and to suggest that our results of lack of a
2 significant correlation between stopping and reward indices were statistically meaningful.
3 For the cross validation analysis, all trials within a neuron were randomly separated to
4 two groups, A and B. Stopping and reward index were computed for those two groups of
5 each neuron. We performed correlations between stopping indices of groups A and B,
6 and between reward indices of A and B. A total of n ($=100$) random permutation
7 instances were performed to generate different A and B sets. The test should ideally show
8 high correlations between indices of A and B for any instance, and we indeed saw
9 positive correlations between stopping-index_A and stopping-index_B, and similarly for
10 reward-index_A and reward-index_B. We confirmed that the actual correlation coefficient
11 between stopping and reward indices in OFC fell within bottom 5% of the coefficients
12 computed for n instances of stopping-index_A and stopping-index_B. The above was also
13 confirmed for n coefficients for reward-index_A and reward-index_B. *Figures 5C-D* showed
14 results of no-significant correlations between stopping and reward indices with $p < 0.01$.

15

16

RESULTS

Subjects showed typical behavior in the stop signal task

Subjects performed a standard stop signal task (based on Hanes and Schall, 1995, *Figure 1A* and Methods). On each trial, following a central fixation, monkeys saw an eccentric target (go signal) that, if fixated, provided a juice reward. On a subset of trials (33%, called stop trials), a second signal (stop signal) appeared at fixation and countermanded the previously instructed saccade. Successful inhibition trials were rewarded. Failed trials (trials in which a saccade was made despite a stop signal) were not. Both subjects showed typical behavior in this task; their performance in stop trials varied as a function of time of presentation of stop signals relative to that of go signal (*Figures 1D, 1G*). Median reaction time in go trials was 0.41 sec and 0.27 sec in subject J and subject T, respectively (*Figures 1C, 1F*).

The delay between the go signal and the stop signal is called the stop signal delay (SSD) and it varied randomly across trials. We estimated the SSD that leads to approximately 50% successful stopping (SSD-50) because it can help in computing the stop signal reaction time (SSRT, Logan, 1994; Logan and Cowan, 1984; Verbruggen and Logan, 2008). The SSD-50 was 0.27 sec for subject J and 0.15 sec for subject T. SSRT was 0.14 sec for subject J, and 0.12 sec for subject T. These values are typical of rhesus macaques in these tasks (e.g. Hanes and Schall, 1995; Ito et al., 2003).

Both subjects showed behavioral effects in the reaction times of successful inhibitions as a function of previous trial conditions (*Figure 1E* for subject J, *Figure 1H*

1 for subject T). Successful inhibition trials were shorter when following a successful
 2 inhibition trial (subject J: N = 328, subject T: N = 357) as opposed to following a failed
 3 inhibition (subject J: N = 111, subject T: N = 60). The statistics for subject J was 360
 4 msec shorter, t-test, t-stat = 11.33, $p < 0.0001$ and for subject T was 290 msec shorter, t-
 5 stat = 11.88, $p < 0.0001$. Similarly, successful inhibition trials were shorter when
 6 following a go trial (subject J: N = 833, subject T: N = 862) as opposed to following a
 7 failed inhibition trial (subject J: 310 msec shorter, t-stat = 9.608, $p < 0.0001$ and subject
 8 T: 210 msec shorter, t-stat = 7.72, $p < 0.0001$).

9

Selectivity for stopping in single neurons

11 We recorded responses of 96 neurons (52 in subject J and 44 in subject T) in area
 12 13 of the OFC (*Figure 1B*). The number of neurons to be collected was determined *a*
 13 *priori* based on exploratory analyses of previous datasets and was not adjusted during
 14 recording based on analyses performed mid-experiment. Responses of example neurons
 15 are illustrated in *Figure 2*. We focus on neural responses throughout the trial to make it
 16 easy to compare the stopping related responses at the time periods before and after SSRT.
 17 The responses shown in *Figures 2A and 2B* are aligned to the go signal (time zero). Note
 18 that while these response patterns are conveniently illustrative, they do not necessarily
 19 stand in for the properties of the entire population (see below).

20 In neuron J19 firing rates following the go signal but before the SSRT were lower
 21 on successfully inhibited trials (1.8 spikes/sec) than on failed inhibition trials (4.1
 22 spikes/sec, Wilcoxon rank test, ranksum = 1480, $p < 0.05$, $n = 567$ trials, *Figure 2A*).
 23 Note that there is a larger and more prominent modulation in firing rate later in the trial.

1 Given its timing, this modulation likely relates to outcome monitoring, is too late to
2 influence stopping, and is not of interest here. Another example neuron, T25, showed
3 distinct patterns for successful and failed inhibition trials even 500 msec before the
4 beginning of the trial (ranksum = 2080, $p < 0.05$, $n = 579$ trials, *Figure 2B*).

5 The responses shown in *Figures 2C and 2D* are aligned to stop signal (time zero).
6 *Figure 2C* illustrates the activity of the same neuron shown in *Figure 2A*; its response
7 pattern showed significant differences between successful inhibition trials (1.8 spikes /
8 sec) and failed inhibition trials (4.4 spikes / sec) that begin after the presentation of stop
9 signal but before SSRT (ranksum = 1340, $p < 0.05$). Finally, neuron T10 (*Figure 2D*)
10 fired more vigorously on successful than on failed inhibition trials at around 100 msec
11 after the SSRT (ranksum = 2229, $p < 0.05$). Simple population analyses suggest that these
12 individual neurons are somewhat atypical, however.

13 Relation between stopping and reward related responses for the sample neurons
14 are dealt in detail in the later part of the manuscript.

15

16 **Population averages provide weak information about stopping**

17 Our central hypothesis was that OFC predictively distinguishes successful from
18 failed inhibition. We focused our analyses on two time periods of the trial: 1) The post-
19 stop signal, but pre-SSRT period, and 2) the pre-go signal time period. The post-stop
20 epoch is important because it is when inhibition generated in response to countermanding
21 commands would presumably occur, and has therefore been the focus of many studies of
22 stopping (Schall, 2001; Schall et al., 2002; Logan et al., 2015). It corresponds to the time
23 during which reactive control occurs (Stuphorn and Emeric, 2012). The pre-go signal

1 epoch corresponds to a time before the trial begins; signal differences here presumably
2 reflect proactive control (Stuphorn and Emeric, 2012).

3 Analysis of single neurons did not provide strong evidence for a role for OFC in
4 stopping. The percent of neurons that individually distinguish successful and failed
5 inhibition trials (regardless of sign) was 8.43% during the 100 msec post-stop signal time
6 period, and was 10.50% during the 100 msec pre-go signal time period. (These epochs
7 were selected before analysis in order to reduce the likelihood of p-hacking). These
8 proportions were not significantly greater than chance in either of the two key epochs
9 (chi-square stat = 1.22, $p = 0.26$ in the post-stop signal time period; chi-square stat = 1.8,
10 $p = 0.17$ in the pre-go signal time period). This lack of a detectable effect does not imply
11 that a correlation between stopping and unit activity in OFC does not exist; rather it
12 suggests that if it does exist it is too weak to detect using conventional methods that focus
13 on single neurons in a sample of the size we collected.

14 We next tested whether successful and failed inhibition trials have a consistent
15 sign of effect on firing rates. The percent of significantly positive cells (successful >
16 failed) was 5.40%, and wasn't significantly different from the percent of significantly
17 negative (successful < failed) cells (3.03% chi-square test, chi-square stat = 0.52, $p =$
18 0.47) in the post-stop signal period. The difference in the sizes of the two cell classes was
19 also not significant before the start of trial at the pre-go signal time period (significantly
20 positive cells 7.55%, significantly negative cells 2.95%, chi square = 2.40, $p = 0.12$).

21 Next we looked at grand averages of populations of neurons (*Figure 3*). We
22 observed no difference between successful and failed inhibition trials either after the stop
23 signal or before the beginning of trial. Specifically, during the post-stop signal time

1 period, responses were slightly less for successful than failed inhibition in subject J
 2 (average of 0.3 spikes/sec, $p = 0.6$, *Figure 3B*); the opposite pattern was observed in
 3 subject T (average of 0.52 spikes/sec, $p = 0.53$, *Figure 3D*). Neither effect was
 4 statistically significant. Thus, these results suggest that conventional population averages
 5 don't reveal information about the pattern of stopping. Together these analyses indicate
 6 that, if stopping correlates exist in OFC, they are of a different form than they take in
 7 regions like FEF and SC.

9 **Ensemble patterns strongly distinguish successful from failed stopping**

10 Many studies suggest that ensemble patterns possess properties to code for neural
 11 information and dynamics, which may not be expressed at the level of single units
 12 (Averbeck et al., 2006; Meyers et al., 2008; Zemel et al., 1998). Some studies extend
 13 these thoughts to suggest that information stored in patterns are more important for neural
 14 processing than that present in single units (Morcos et al., 2018). Taking inspiration from
 15 such recent developments in the theoretical understanding of neural activity, we devised
 16 our next analysis on ensemble patterns. We trained neural network decoders to analyze
 17 differences in population activation patterns between successful and failed inhibition that
 18 weren't measured through unit responses or population averages. We were, again,
 19 interested in two time periods: 1) the times after the presentation of the stop signal (which
 20 we examined using a decoder trained on post-stop signal pattern, referred to below as
 21 post-stop decoder), and 2) the time before the start of trial (which we examined using a
 22 decoder trained on pre-go signal pattern, referred to below as pre-go decoder). To ensure
 23 we had enough data to detect significant effects we used 100 msec moving boxcars and to

gain some insight into the time course of effects, we used a 10 msec step size for boxcars.

The post stop signal decoder was able to classify success of an inhibition significantly in a series of 9 consecutive boxcar bins spanning 40 msec after the stop signal to 120 msec after it (these times indicate the starts of the 100 msec boxcars). The numbers for individual subjects were 40 to 140 msec in subject J and 40 to 220 msec in subject T; the central points of the first of these bins is 90 msec for both subjects. These series are unlikely to occur by chance ($p < 0.001$ in all cases, see Methods for specific use of chi-square tests to quantify significance of consecutive bins, and *Figure S1*). Notably, the central point of the first bin of the series to reach significance in both subjects occurred before the stop signal reaction time of either subjects (the SSRTs were 140 msec for subject J and 120 msec, also see *Figure S1*; *Figure 4B*). We call the central point the cancellation time; it measures the center point latency of first statistically significant difference between successful and failed inhibition trials for the ensemble of neurons. The cancellation time is 90 msec for both subjects. The cancellation time preceded the average stopping response by 50 msec in subject J, and by 30 msec in subject T, suggesting OFC's responses may precede the stopping response (see Discussion, and *Figure S1*).

We then examined the response differences of the pre-go signal decoder. We observed a significant pattern difference between successful and failed inhibition trials ($p < 0.001$) extending from 470 to 120 msec before the go signal (see Methods on procedures to determine statistical significance of a boxcar using chi-square statistics). For subject J, significant decoding was observed during the time periods 460 to 120 msec; for subject T it was 420 to 200 msec. These results indicate that the upcoming

1 success or failure of inhibition is decodable from OFC patterns even before the start of
 2 the trial (*Figure 4A*, also see *Figure S1*); Our results do not tell us why this correlation
 3 exists, although one may infer that it reflects some internal state facilitated by variety of
 4 factors such as frequency of task events, frequency of different trial types, motivation,
 5 trial sequence, altogether driving successful versus failed inhibition (Chen, Stuphorn, J
 6 Neuro, 2010); thus it is a likely correlate of proactive control.

8 **The post-stop and pre-go decoders are statistically orthogonal**

9 We next examined how the two decoders related to each other. That is, we asked
 10 whether the patterns that distinguish successful and failed inhibition after the stop signal
 11 related to those that predict inhibition before the trial begins? We did so by comparing
 12 the vector of weights of the post-stop decoder and pre-go decoder. We found a very low
 13 similarity between them (similarity coefficient, ' r ', obtained at zero lag on cross
 14 correlating weight patterns, $r = 0.02 \pm 0.22$). This low correlation may be due to noise in
 15 our signal. We therefore performed a cross-validation procedure to measure range of
 16 values expected from a true correlation with noise. The measured coefficient fell below
 17 1st percentile of that obtained from autocorrelation (and thus is significant at $p \leq 0.01$;
 18 100 randomizations, average r from the randomized sets = 25.84 ± 0.82). The decoding
 19 performances during the time periods after the stop signal (t-stat = 6.0491, $p = 0.003$) and
 20 before the beginning of trial (t-test, t-stat = 8.8874, $p < 0.001$) were significantly different
 21 between the networks trained on pre-go and post-stop signals. The above results suggest
 22 that the two decoders that predict successful vs. failed inhibition are statistically
 23 orthogonal and thus dissimilar in the early and late epochs.

1

2 **Activity of OFC ensembles, but not single neurons, correlates with reaction time**

3 Neurons in prefrontal structures such as SEF show linear correlations between
4 single neural responses and trial reaction time (Stuphorn et al., 2010). We next asked
5 whether single neuron responses from our OFC data around response showed correlations
6 to reaction time, the time taken to saccade to the choice target. We computed reaction
7 time as the time difference between the presentation of go signal and time of saccade to
8 target (on trials with such saccades, go trials and failed inhibition trials, Hanes et al.,
9 1996; Stuphorn et al., 2010).

10 We used mean firing rates during 200 msec before and after the reaction time for
11 this analysis; our analytical approach was designed to be similar to that used by Stuphorn
12 et al. (2010). We found no correlations between reaction time and firing rates before the
13 reaction time (Pearson correlation, $\rho = 0$, $p = 0.41$). Likewise, we found no correlation
14 after the reaction time (Pearson correlation, $\rho = -0.01$, $p = 0.09$). This analysis suggests
15 that activity of single neurons in OFC, unlike those in SEF, doesn't scale linearly with
16 reaction time (Stuphorn et al., 2010). This lack of observed correlation raises the
17 possibility that downstream regions may not linearly decode the information from OFC
18 for informing the urgency of action execution.

19 We hypothesized that OFC ensemble responses predict reaction times. To test this
20 idea, we generated population activation patterns from neurons that contained data for
21 discretized reaction time bins in a range of 0.1 sec to 0.6 sec with step size 100 msec (5
22 equally sized bins, see Methods for the specific criterion used). In particular, we asked
23 whether the ensemble response could be accurately classified to discrete reaction time

bins (see Methods) in a non-linear fashion. The results show that neural network decoders were able to classify OFC ensembles to correct reaction time bins (See Methods, t -test, t -stat = 27.91, $p < 0.001$), when the population activation pattern was generated from 200 msec time periods before and after the reaction time (chi-square test, $p < 0.05$ for all 5 bins). These results suggest that OFC ensemble responses can predict reaction times.

6

7 **OFC codes for stopping and reward are unrelated**

8 The reward-encoding role of OFC is a hallmark of its function (Padoa-Schioppa, 9 2011; Schultz, 2000; Wallis, 2007). We therefore wondered whether the stopping-related 10 activity that we observed might be a side-effect of its reward roles. For example, it may 11 be that there is some undetectable natural variation in the relative subjective value of the 12 reward offered for correct performance. On trials in which the reward happened to have a 13 slightly lower value, the subject would be less motivated to perform correctly; this 14 fluctuation would then introduce a correlation between firing rates and successful 15 inhibition.

16 To test for the possibility that our putative inhibition signals were just reward 17 correlates, we took advantage of a second set of trials, collected in a *neuroeconomic* 18 *stopping task*; detailed analysis of the results from that task will be the focus of a later 19 manuscript. In this task, subjects chose or rejected a single reward that had one of three 20 values (low, medium, and high rewards, see Methods). The two task types, 21 neuroeconomic and standard stop signal paradigms, were randomly interleaved on a trial- 22 by-trial basis. The data from this task allowed us to assess each neuron's tuning function 23 for anticipated rewards. Responses to different reward amounts by two example neurons 24 are shown in *Figures 5A and 5B*. We found tuning for anticipated reward values in the

1 firing activity during the reward feedback time period. For example, we observed a
 2 significant positive correlation between reward amount and firing rate in neuron J19 ($\rho =$
 3 $0.3138, p < 0.001$, *Figure 5A, same as Figure 2A but aligned to feedback*) and a
 4 significant negative one in neuron T10 ($\rho = -0.143, p = 0.04$, *Figure 5B*).

5 If the stopping-related signals were a consequence of reward encoding, we would
 6 see a positive correlation between coding patterns for rewards and stopping. We
 7 computed a *reward index* for all neurons by regressing their responses to outcomes
 8 against the outcomes themselves. We computed a stopping index for all neurons by
 9 subtracting on their firing rate during successful and failed inhibition before the stop
 10 signal reaction time (see Methods). We found no correlations between these indices in the
 11 post-stop signal time period (Pearson correlation, $\rho = 0.09, p = 0.4$, *Figure 5D*). Nor did we
 12 find such correlations in pre-go signal time period ($\rho = -0.02, p = 0.82$, *Figure 5C*).

13 This lack of correlation may be a sign that the reward code and the stopping code
 14 are different. It may also, in theory, be due to lack of sufficient data to detect a significant
 15 effect. To test this idea, we performed a cross-validation analysis (See Methods).
 16 Specifically, we reasoned that if insufficient data were the problem then a within sample
 17 correlation would also produce no significant correlation. A positive correlation of a
 18 within sample correlation, using randomly sampled half-sized subsets, then, would
 19 indicate that our data have sufficient power to detect a significant effect (Blanchard et al.,
 20 2015). We thus tested whether the correlation coefficient for stopping and reward indices
 21 fell below the bottom 5 percentile of the coefficients obtained for within-group
 22 correlations. Indeed, the coefficient fell below 1st percentile of that obtained for 100

1 randomizations in cross validation analysis. *Figures 5C-D* show no-correlations between
2 stopping and reward indices with $p \leq 0.01$.
3
4

DISCUSSION

We examined the correlates of successful stopping in ensembles of neurons in OFC. We found that while traditional single unit correlates of stopping associated with more dorsal structures are too weak to observe (if there at all), that ensembles of neurons in OFC readily distinguish successful and failed inhibition. These signals were not consistently associated with a higher or lower firing rate, nor were they associated with two discrete sets of neurons, as in FEF and SC (Hanes et al., 1998; Pouget et al., 2017; Stuphorn et al., 2000). These findings, then, raise the possibility that OFC plays a somewhat different role in regulation of stopping than these other regions.

Our study shows the presence of stopping-related patterns in OFC at two specific time periods, the first one after the presentation of stop signal, and the second one before the beginning of trial. The timing of the two stopping-related patterns is reminiscent of the times associated with reactive and proactive control, respectively (Braver, 2012; Braver et al., 2007; Chen et al., 2010; Chikazoe et al., 2009; Hanes et al., 1998; Ito et al., 2003; Majid et al., 2013; Stuphorn et al., 2000; Stuphorn et al., 2010; Stuphorn and Emeric, 2012). These response patterns indicate that OFC neurons carry a signal that precedes stopping decisions. They suggest, then, that OFC may be part of the pathway that determines the success or failure of stopping.

OFC is marked by its unusual anatomy: it receives strong and diverse sensory inputs, as well as visceral ones, and it projects to more dorsal prefrontal structures that collectively directly regulate behavior (Ongur and Price, 2000; Cavada et al., 2000, Wallis, 2007). Its connections then mean that it has a nearly ideal anatomy for monitoring

1 sensory and reward information forming a first draft of the type of executive signals that
 2 can inform - but not determine - action. Its influence is unlikely to be limited to
 3 inhibition; its executive functions likely include contingent (rule-based) decisions,
 4 working memory, switching, and conflict monitoring (Bryden and Roesch, 2015; Lara et
 5 al., 2009; Mansouri et al., 2014; Sleezer et al., 2016; Sleezer et al., 2017). More broadly,
 6 these executive signals likely constitute a component of a larger set of output prediction
 7 signals that determine OFC's major role in cognition (Rudebeck and Murray, 2014;
 8 Schuck et al., 2016; Wilson et al., 2014).

9 Another well-known role of the OFC is in signaling value (Padoa-Schioppa, 2011;
 10 Wallis, 2007). Given this close association, one might expect that its role in inhibition is
 11 to signal the current subjective value of stopping. Our data, which indicate that OFC
 12 value responses are orthogonal to its stopping responses, argue against this possibility.
 13 They are consistent, however, with the broader theory that OFC carries a suite of signals
 14 that regulate the ongoing transformation of stimulus into action, and that value is one
 15 such signal. Other executive signals observed in OFC include spatial information and
 16 abstract rules (Lara et al., 2009; Luk and Wallis, 2013; Sleezer et al., 2016; Strait et al.,
 17 2016; Tsujimoto et al., 2009; Wallis, 2007; Wallis et al., 2001), those related to switching
 18 and task adjustment (Chase et al., 2012; Sleezer et al., 2017), and to conflict adaptation
 19 (Bryden and Roesch, 2015; Mansouri et al., 2014; Mansouri et al., 2009).

20 In foraging theory, decisions are generally framed as accept-reject. From this
 21 perspective, binary choices, the mainstay of behavioral economics and microeconomics,
 22 are actually better thought of as two somewhat independent accept-reject decisions
 23 (Kacelnik et al., 2011). Each accept-reject decision, in turn, functions like a classic

1 accept-reject foraging decision, that is, as a choice between pursuing and refraining from
 2 pursuit. In other words, what appears to be a binary choice may actually be a pair of
 3 stopping decisions. If economic choice ultimately boils down to stopping, there is an
 4 opportunity for a “grand unified theory” that can explain the two types of decisions. This
 5 possibility would help explain, for example, why many of the same regions are involved
 6 in both types of decisions. In particular, OFC has demonstrated importance in both
 7 economic decisions and in inhibition. Progress in this area promises to help shed light on
 8 important debates, such as how economic choice relates to self-control (Berkman et al.,
 9 2016; Shenhav, 2017).

10

11

12

13

1

2 **Acknowledgements:**

3

4 This work was supported by an R01 (DA038615) to BYH. We thank Meghan C. Pesce
5 for help with data collection.

6

7 The authors declare no conflicts of interest.

8

FIGURES

Figure 1- stop signal task and subject behavior: (a) task framework (b) recording site - Area 13 of OFC (scan from subject J shown) Behavioral results for subject J are presented in panels (c, d, e) and for subject T in (f, g, h) (c, f) Go-trial reaction time distributions (d, g) Inhibition function varied as a function of SSDs (e, h) Previous trial had effects in reaction time behavior. Red lines in (c, e) denote the summation SSD-50+ SSRT (Logan 1984; Logan and Cowan, 1984), and red-lines in (d, g) denote SSD-50. Error bars in (e, h) represent SEM, and * denotes t-test significance with $p < 0.05$.

Figure 2- selectivity for stopping in sample neurons: Activity of example neurons during successful inhibition, failed inhibition and go trials are illustrated with respect to (a, b) go signal (c, d) stop signal presentation time. Time from start of the go (stop) signal to SSRT is shaded in panels A and B (C and D). Neuron in panel A shows significant difference in firing rates of successful and failed inhibition trials before SSRT. Neuron in panel B shows difference even before the beginning of trial. Neuron in panel C is the same as panel A, shows significant difference in firing rates of inhibition trials before stopping response time. Likewise, neuron in panel D shows difference around few msec after SSRT.

Figure 3- population averages provide weak information about stopping: Population activity for successful inhibition and failed inhibition with respect to (a, c) go signal presentation and (b, d) stop signal presentation, for subjects J and T. Time from start of the go (stop) signal to SSRT is shaded in panels A and C (B and D). Data for all SSDs are averaged to present successful and failed inhibition trials. Error bars denote SEM. They don't reveal significant information about the pattern of stopping.

Figure 4- ensemble analysis inform about stopping: Performance of (a) pre-go signal (b) post-stop decoders to distinguish successful vs. failed inhibition pattern. Insets in both panels illustrate projections of decoder's output responses (o_1 and o_2 , for successful and failed inhibition trials, respectively). Error bars represent SEM. Time points highlighted through horizontal bars denote start time of 100 msec boxcars having percent accuracies of classification above chance of 50% (chi-square test, $p < 0.05$).

Figure 5- unrelated reward and stopping codes: (a, b) Illustration of example neurons tuned to reward sizes. Neuron in panel A (panel B) shows significant positive (negative) correlation to reward amounts. Correlations between stopping

1 and reward indices show no significant effect during 100 msec in **(c)** pre-go
2 signal and **(d)** post-stop signal time period.
3
4

FIGURES

Figure 1

1 **Figure 2**

2

1 **Figure 3**

2
3

1 **Figure 4**

2

3

1 **Figure 5**

2

3

SUPPLEMENTARY MATERIAL

Figure S1- Post-stop and pre-go decoding for subjects J and T: The post stop signal decoder was able to classify success of an inhibition significantly above chance (see Methods for specific use of chi-square tests to quantify significance) in a time period ranging from 40 msec to 170 msec for subject J, and 40 to 220 msec for subject T, respectively after the stop signal (these times indicate the beginning of 100 msec boxcars, and chi-square tests were used for finding their significance with $p < 0.05$, see Methods). The first significant bin was therefore of window size 40 – 140 msec, that led to average cancellation time as 90 msec. It preceded the average stopping response by 50 msec in subject J, and by 30 msec in subject T, suggesting OFC's responses may precede the stopping response. In Pre-go signal decoder, for subject J, high accuracy of decoding was found during the time periods 460 msec to 120 msec before the appearance of go signal. Likewise, it was 420 msec to 200 msec in subject T.

Figure S2- Post-stop and pre-go decoding for subjects J and T with zscored data (normalization): Both post-stop signal and pre-go signal decoder was able to classify success of stopping significantly above chance (see **Methods** for specific use of chi-square tests to quantify significance) before SSRT and go signal presentation, respectively, and chi-square tests were used for finding their significance with $p < 0.05$, see **Methods**). Results suggest that successful differentiation of stopping codes can be obtained irrespective of the normalization methods used in the study (In the manuscript, Normalization procedure was carried out by subtracting the mean firing during inter-trial interval (ITI) time period (baseline) and then by zscoring each neuron's data, and the normalized data is used for decoder analysis).

REFERENCES

- Aron, A.R. (2007). The neural basis of inhibition in cognitive control. *The neuroscientist* 13, 214-228.
- Aron, A.R., Cai, W., Badre, D., and Robbins, T.W. (2015). Evidence supports specific braking function for inferior PFC. *Trends Cogn Sci* 19, 711-712.
- Aron, A.R., Fletcher, P.C., Bullmore, E.T., Sahakian, B.J., and Robbins, T.W. (2003). Stop-signal inhibition disrupted by damage to right inferior frontal gyrus in humans. *Nature neuroscience* 6, 115-116.
- Averbeck, B.B., Latham, P.E., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nature reviews neuroscience* 7, 358.
- Banich, M.T., and Depue, B.E. (2015). Recent advances in understanding neural systems that support inhibitory control. *Current opinion in behavioral sciences* 1, 17-22.
- Berkman, E., Hutcherson, C., Livingston, J.L., Kahn, L.E., and Inzlicht, M. (2016). Self-control as value-based choice.
- Blanchard, T.C., Hayden, B.Y., and Bromberg-Martin, E.S. (2015). Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. *Neuron* 85, 602-614.
- Brainard, D.H., and Vision, S. (1997). The psychophysics toolbox. *Spatial vision* 10, 433-436.
- Braver, T.S. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in cognitive sciences* 16, 106-113.
- Braver, T.S., Gray, J.R., and Burgess, G.C. (2007). Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. *Variation in working memory*, 76-106.
- Bryden, D.W., and Roesch, M.R. (2015). Executive control signals in orbitofrontal cortex during response inhibition. *Journal of Neuroscience* 35, 3903-3914.
- Carmichael, S., and Price, J. (1994). Architectonic subdivision of the orbital and medial prefrontal cortex in the macaque monkey. *Journal of Comparative Neurology* 346, 366-402.
- Carmichael, S., and Price, J. (1996). Connectional networks within the orbital and medial prefrontal cortex of macaque monkeys. *Journal of Comparative Neurology* 371, 179-207.
- Chamberlain, S.R., Fineberg, N.A., Blackwell, A.D., Robbins, T.W., and Sahakian, B.J. (2006). Motor inhibition and cognitive flexibility in obsessive-compulsive disorder and trichotillomania. *American Journal of Psychiatry* 163, 1282-1284.
- Chase, E.A., Tait, D.S., and Brown, V.J. (2012). Lesions of the orbital prefrontal cortex impair the formation of attentional set in rats. *European Journal of Neuroscience* 36, 2368-2375.
- Chen, X., Scangos, K.W., and Stuphorn, V. (2010). Supplementary motor area exerts proactive and reactive control of arm movements. *Journal of Neuroscience* 30, 14657-14675.
- Chikazoe, J., Jimura, K., Hirose, S., Yamashita, K.-i., Miyashita, Y., and Konishi, S. (2009). Preparation to inhibit a response complements response inhibition during performance of a stop signal task. *Journal of Neuroscience* 29, 15870-15877.

- 1 Chudasama, Y., Kralik, J., and Murray, E. (2006). Rhesus monkeys with orbital
2 prefrontal cortex lesions can learn to inhibit prepotent responses in the reversed reward
3 contingency task. *Cerebral Cortex* 17, 1154-1159.
- 4 Cornelissen, F.W., Peters, E.M., and Palmer, J. (2002). The Eyelink Toolbox: eye
5 tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods*
6 34, 613-617.
- 7 Dias, R., Robbins, T., and Roberts, A. (1996). Dissociation in prefrontal cortex of
8 affective and attentional shifts. *Nature* 380, 69-72.
- 9 Duncan, J. (2001). An adaptive coding model of neural function in prefrontal
10 cortex. *Nature Reviews Neuroscience* 2, 820-829.
- 11 Eagle, D.M., Bari, A., and Robbins, T.W. (2008). The neuropsychopharmacology
12 of action inhibition: cross-species translation of the stop signal and go/no-go tasks.
13 *Psychopharmacology* 199, 439-456.
- 14 Eagle, D.M., Baunez, C., Hutcheson, D.M., Lehmann, O., Shah, A.P., and
15 Robbins, T.W. (2007). Stop-signal reaction-time task performance: role of prefrontal
16 cortex and subthalamic nucleus. *Cerebral cortex* 18, 178-188.
- 17 Fuster, J.M. (1988). Prefrontal cortex. In *Comparative neuroscience and*
18 *neurobiology* (Springer), pp. 107-109.
- 19 Fuster, J.n.M. (2001). The prefrontal cortex—an update: time is of the essence.
20 *Neuron* 30, 319-333.
- 21 Ghods-Sharifi, S., Haluk, D.M., and Floresco, S.B. (2008). Differential effects of
22 inactivation of the orbitofrontal cortex on strategy set-shifting and reversal learning.
23 *Neurobiology of learning and memory* 89, 567-573.
- 24 Hampshire, A., and Sharp, D.J. (2015). Contrasting network and modular
25 perspectives on inhibitory control. *Trends in cognitive sciences* 19, 445-452.
- 26 Hanes, D.P., and Carpenter, R. (1999). Countermanding saccades in humans.
27 *Vision research* 39, 2777-2791.
- 28 Hanes, D.P., Patterson, W.F., and Schall, J.D. (1998). Role of frontal eye fields in
29 countermanding saccades: visual, movement, and fixation activity. *Journal of*
30 *Neurophysiology* 79, 817-834.
- 31 Hanes, D.P., and Schall, J.D. (1995). Countermanding saccades in macaque.
32 *Visual neuroscience* 12, 929-937.
- 33 Hanes, D.P., and Schall, J.D. (1996). Neural control of voluntary movement initiation.
34 *Science* 274, 427.
- 35 Haykin, S., and Network, N. (2004). A comprehensive foundation. *Neural*
36 *Networks* 2, 41.
- 37 Horn, N., Dolan, M., Elliott, R., Deakin, J., and Woodruff, P. (2003). Response
38 inhibition and impulsivity: an fMRI study. *Neuropsychologia* 41, 1959-1966.
- 39 Iacono, W.G., Malone, S.M., and McGue, M. (2008). Behavioral disinhibition and
40 the development of early-onset addiction: common and specific influences. *Annu Rev*
41 *Clin Psychol* 4, 325-348.
- 42 Ito, S., Stuphorn, V., Brown, J.W., and Schall, J.D. (2003). Performance
43 monitoring by the anterior cingulate cortex during saccade countermanding. *Science*
44 302, 120-122.
- 45 Kacelnik, A., Vasconcelos, M., Monteiro, T., and Aw, J. (2011). Darwin's "tug-of-
46 war" vs. starlings' "horse-racing": how adaptations for sequential encounters drive
47 simultaneous choice. *Behavioral Ecology and Sociobiology* 65, 547-558.
- 48 Lara, A.H., Kennerley, S.W., and Wallis, J.D. (2009). Encoding of gustatory
49 working memory by orbitofrontal neurons. *Journal of Neuroscience* 29, 765-774.

- 1 Li, C.-s.R., Huang, C., Constable, R.T., and Sinha, R. (2006). Imaging response
2 inhibition in a stop signal task: neural correlates independent of signal monitoring and
3 post-response processing. *Journal of Neuroscience* 26, 186-192.
- 4 Logan, G.D. (1994). On the ability to inhibit thought and action: A users' guide to
5 the stop signal paradigm.
- 6 Logan, G.D., and Cowan, W.B. (1984). On the ability to inhibit thought and
7 action: A theory of an act of control. *Psychological review* 91, 295.
- 8 Logan, G.D., and Irwin, D.E. (2000). Don't look! Don't touch! Inhibitory control of
9 eye and hand movements. *Psychonomic Bulletin & Review* 7, 107-112.
- 10 Logan, G.D., Schachar, R.J., and Tannock, R. (1997). Impulsivity and inhibitory
11 control. *Psychological science* 8, 60-64.
- 12 Logan, G.D., Yamaguchi, M., Schall, J.D., and Palmeri, T.J. (2015). Inhibitory
13 control in mind and brain 2.0: blocked-input models of saccadic countermanding.
14 *Psychological review* 122, 115.
- 15 Luk, C.-H., and Wallis, J.D. (2013). Choice coding in frontal cortex during
16 stimulus-guided or action-guided decision-making. *The Journal of Neuroscience* 33,
17 1864-1871.
- 18 MacLeod, C.M., Dodd, M.D., Sheard, E.D., Wilson, D.E., and Bibi, U. (2003). In
19 opposition to inhibition. *Psychology of learning and motivation* 43, 163-215.
- 20 Majid, D.A., Cai, W., Corey-Bloom, J., and Aron, A.R. (2013). Proactive selective
21 response suppression is implemented via the basal ganglia. *Journal of Neuroscience*
22 33, 13259-13269.
- 23 Mansouri, F.A., Buckley, M.J., and Tanaka, K. (2014). The essential role of
24 primate orbitofrontal cortex in conflict-induced executive control adjustment. *Journal of*
25 *Neuroscience* 34, 11016-11031.
- 26 Mansouri, F.A., Tanaka, K., and Buckley, M.J. (2009). Conflict-induced
27 behavioural adjustment: a clue to the executive functions of the prefrontal cortex.
28 *Nature reviews Neuroscience* 10, 141.
- 29 Mante, V., Sussillo, D., Shenoy, K.V., and Newsome, W.T. (2013). Context-
30 dependent computation by recurrent dynamics in prefrontal cortex. *nature* 503, 78.
- 31 Meyers, E.M., Freedman, D.J., Kreiman, G., Miller, E.K., and Poggio, T. (2008).
32 Dynamic population coding of category information in inferior temporal and prefrontal
33 cortex. *Journal of neurophysiology* 100, 1407-1419.
- 34 Mishkin, M. (1964). Perseveration of central sets after frontal lesions in monkeys.
35 *The frontal granular cortex and behavior*, 219-241.
- 36 Morcos, A.S., Barrett, D.G., Rabinowitz, N.C., and Botvinick, M. (2018). On the
37 importance of single directions for generalization. *arXiv preprint arXiv:180306959*.
- 38 Nederkoorn, C., Braet, C., Van Eijs, Y., Tanghe, A., and Jansen, A. (2006). Why
39 obese children cannot resist food: the role of impulsivity. *Eating behaviors* 7, 315-322.
- 40 Öngür, D., and Price, J. (2000). The organization of networks within the orbital
41 and medial prefrontal cortex of rats, monkeys and humans. *Cerebral cortex* 10, 206-
42 219.
- 43 Padoa-Schioppa, C. (2011). Neurobiology of economic choice: a good-based
44 model. *Annual review of neuroscience* 34, 333.
- 45 Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with
46 population codes. *Nature reviews Neuroscience* 1, 125.
- 47 Pouget, P., Murthy, A., and Stuphorn, V. (2017). Cortical control and
48 performance monitoring of interrupting and redirecting movements. *Phil Trans R Soc B*
49 372, 20160201.

- 1 Rigotti, M., Barak, O., Warden, M.R., Wang, X.-J., Daw, N.D., Miller, E.K., and
2 Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*
3 497, 585.
- 4 Roberts, A., and Wallis, J. (2000). Inhibitory control and affective processing in
5 the prefrontal cortex: neuropsychological studies in the common marmoset. *Cerebral*
6 *Cortex* 10, 252-262.
- 7 Rudebeck, P.H., and Murray, E.A. (2014). The orbitofrontal oracle: cortical
8 mechanisms for the prediction and evaluation of specific behavioral outcomes. *Neuron*
9 84, 1143-1156.
- 10 Rumelhart, D., McClelland, J., and Williams, R. (1986). Parallel recognition in
11 modern computers. processing: Explorations in the microstructure of cognition 1.
- 12 Rushworth, M.F., Kolling, N., Sallet, J., and Mars, R.B. (2012). Valuation and
13 decision-making in frontal cortex: one or many serial or parallel systems? *Current*
14 *opinion in neurobiology* 22, 946-955.
- 15 Rushworth, M.F., Noonan, M.P., Boorman, E.D., Walton, M.E., and Behrens, T.E.
16 (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron* 70,
17 1054-1069.
- 18 Schachar, R., Tannock, R., Marriott, M., and Logan, G. (1995). Deficient
19 inhibitory control in attention deficit hyperactivity disorder. *Journal of abnormal child*
20 *psychology* 23, 411-437.
- 21 Schall, J.D. (1991). Neuronal activity related to visually guided saccades in the
22 frontal eye fields of rhesus monkeys: comparison with supplementary eye fields.
23 *Journal of neurophysiology* 66, 559-579.
- 24 Schall, J.D. (2001). Neural basis of deciding, choosing and acting. *Nature*
25 *Reviews Neuroscience* 2, 33-42.
- 26 Schall, J.D., Stuphorn, V., and Brown, J.W. (2002). Monitoring and control of
27 action by the frontal lobes. *Neuron* 36, 309-322.
- 28 Schoenbaum, G., Roesch, M.R., Stalnaker, T.A., and Takahashi, Y.K. (2009). A
29 new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nature*
30 *Reviews Neuroscience* 10, 885-892.
- 31 Schoenbaum, G., Setlow, B., Nugent, S.L., Saddoris, M.P., and Gallagher, M.
32 (2003). Lesions of orbitofrontal cortex and basolateral amygdala complex disrupt
33 acquisition of odor-guided discriminations and reversals. *Learning & Memory* 10, 129-
34 140.
- 35 Schuck, N.W., Cai, M.B., Wilson, R.C., and Niv, Y. (2016). Human orbitofrontal
36 cortex represents a cognitive map of state space. *Neuron* 91, 1402-1412.
- 37 Schultz, W. (2000). Multiple reward signals in the brain. *Nature reviews*
38 *Neuroscience* 1, 199
- 39 Shenhav, A. (2017). The perils of losing control: Why self-control is not just
40 another value-based decision. *Psychological Inquiry*
- 41 Sleezer, B.J., Castagno, M.D., and Hayden, B.Y. (2016). Rule encoding in
42 orbitofrontal cortex and striatum guides selection. *Journal of Neuroscience* 36, 11223-
43 11237.
- 44 Sleezer, B.J., LoConte, G.A., Castagno, M.D., and Hayden, B.Y. (2017).
45 Neuronal responses support a role for orbitofrontal cortex in cognitive set
46 reconfiguration. *European Journal of Neuroscience* 45, 940-951.
- 47 Strait, C.E., Sleezer, B.J., Blanchard, T.C., Azab, H., Castagno, M.D., and
48 Hayden, B.Y. (2016). Neuronal selectivity for spatial positions of offers and choices in
49 five reward regions. *Journal of neurophysiology* 115, 1098-1111.

- 1 Stuphorn, V., Bauswein, E., and Hoffmann, K.-P. (2000). Neurons in the primate
2 superior colliculus coding for arm movements in gaze-related coordinates. *Journal of*
3 *Neurophysiology* 83, 1283-1299.
- 4 Stuphorn, V., Brown, J.W., and Schall, J.D. (2010). Role of supplementary eye
5 field in saccade initiation: executive, not direct, control. *Journal of neurophysiology* 103,
6 801-816.
- 7 Stuphorn, V., and Emeric, E.E. (2012). Proactive and reactive control by the
8 medial frontal cortex. *Frontiers in Neuroengineering* 5.
- 9 Tsujimoto, S., Genovesio, A., and Wise, S.P. (2009). Monkey orbitofrontal cortex
10 encodes response choices near feedback time. *Journal of Neuroscience* 29, 2569-
11 2574.
- 12 Verbruggen, F., and Logan, G.D. (2008). Response inhibition in the stop signal
13 paradigm. *Trends in cognitive sciences* 12, 418-424.
- 14 Wallis, J.D. (2007). Orbitofrontal cortex and its contribution to decision-making.
15 *Annu Rev Neurosci* 30, 31-56.
- 16 Wallis, J.D., Anderson, K.C., and Miller, E.K. (2001). Single neurons in prefrontal
17 cortex encode abstract rules. *Nature* 411, 953-956.
- 18 Wang, M.Z., and Hayden, B.Y. (2017). Reactivation of associative structure
19 specific outcome responses during prospective evaluation in reward-based choices.
20 *Nature Communications* 8, ncomms15821.
- 21 Werbos, P. (1974). Beyond regression: New tools for prediction and analysis in
22 the behavioral sciences.
- 23 Wikenheiser, A.M., and Schoenbaum, G. (2016). Over the river, through the
24 woods: cognitive maps in the hippocampus and orbitofrontal cortex. *Nature Reviews*
25 *Neuroscience* 17, 513-523.
- 26 Wilson, R.C., Takahashi, Y.K., Schoenbaum, G., and Niv, Y. (2014). Orbitofrontal
27 cortex as a cognitive map of task space. *Neuron* 81, 267-279.
- 28 Zemel, R.S., Dayan, P., and Pouget, A. (1998). Probabilistic interpretation of
29 population codes. *Neural computation* 10, 403-430
- 30